A STUDY OF THE RELATIONSHIP BETWEEN GDP, POPULATION & CO2 EMISSIONS BASED ON TIME SERIES

CHETAN R TIPPA

Thesis Report

DECEMBER 2023

DEDICATION

To my unwavering source of inspiration, my family, and loved ones,

This thesis stands as a testament to the support, love, and encouragement that you have showered upon me throughout this arduous journey. Your unwavering belief in my abilities and your sacrifices have been the driving force behind my pursuit of knowledge.

To my parents, whose sacrifices and unwavering faith in me have shaped me into the person I am today. Your enduring love and endless encouragement have been my anchor in the stormy sea of academia.

To my siblings, friends, and mentors, who have provided invaluable guidance, encouragement, and camaraderie. Your wisdom and friendship have illuminated my path, making this journey more meaningful.

To my professors and advisors, who have patiently nurtured my intellectual growth and challenged me to reach greater heights. Your dedication to my education has been instrumental in shaping the ideas presented in this thesis.

To my wife who believed in me, even when I doubted myself, and to those who cheered me on during the late-night study sessions and moments of self-doubt, this work is dedicated to you.

May this thesis stand as a tribute to the collective effort and unwavering support that has made it possible. Thank you for being my guiding star on this academic odyssey.

With all my love,

Chetan.R. Tippa

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to several individuals who have played pivotal roles in the completion of my thesis.

First and foremost, I extend my deepest thanks to my thesis mentor, Mr. Karthik O S, whose unwavering guidance, patience, and expertise have been instrumental in shaping this research. Your mentorship has not only broadened my horizons but has also instilled in me a passion for continuous learning.

To my parents, your unceasing support and encouragement have been the bedrock of my academic journey. Your sacrifices and belief in my potential have been my driving force.

To my friends, who have been a constant source of inspiration and motivation, I am grateful for the countless discussions, brainstorming sessions, and the laughter that we shared during this journey.

Last but not least, to my beloved wife, your boundless love, understanding, and unwavering support have been my anchor. You've stood by me through thick and thin, and this thesis is as much a testament to your resilience as it is to mine.

To all of you, I owe my deepest gratitude for being my pillars of strength and for making this achievement possible.

With profound appreciation,

Chetan.R. Tippa

ABSTRACT

China, the US, and India are the largest emitters of CO2 emissions in the world. Motivated by the commitment of India at COP26, we study the relationship between GDP, population, and CO2 emissions of India. This study aims to investigate the link between economic development and environmental deterioration. The study uses the “Our World in Data” dataset on India's GDP, population, and CO2 emissions including changes in land use from 1850 to 2021. We employ both the VAR model & Multiple linear regression model. The study's scope is restricted to comparing the VAR model with the Multiple linear regression model, to identify the most accurate model for the study through model evaluation. The Granger causality test is conducted to ascertain the relationship between variables prior to constructing the model. By performing Granger’s causality test both unidirectional and bidirectional causality can be found between the variables. The variable co2_including_luc_ Granger causes gdp. The VAR model is evaluated using accuracy metrics RMSE and multiple linear regression using R^2. A projection of co2_including_luc, gdp, and population of actual values versus predicted values is performed for the ten-year period from 2011 to 2021, utilizing historical data spanning from 1850 to 2021. The accuracy of the multiple linear regression (MLR) model is found to be higher when compared to the vector autoregression (VAR) model. The multiple linear regression (MLR) model demonstrates a coefficient of determination (R2) of 78%. Furthermore, it is worth noting that the variables gdp, population, and co2_including_luc exhibit root mean square error (RMSE) values of 5.80E+12, 1.2 billion, and 1683 million tonnes, respectively. The projected values for the variables ‘gdp’, ‘population’, and ‘co2_including_luc’ demonstrate a modest similarity and an ascending pattern when compared to the observed data. The findings derived from the study will provide valuable insights for the government to inform and guide potential policy modifications.

Keywords: India, gdp, population, CO2 emissions including changes in land use, VAR, Multiple Linear regression, RMSE, R^2, Granger’s causality test, model accuracy.

LIST OF FIGURES

Figure 1. GHG emissions country-wise percentage from 1990 - 2020 7
Figure 1.1 CO2 emissions country-wise percentage from 1990 - 2020 7
Figure.1.2 Global CO2 emissions from fossil fuels and land use change 8
Figure 1.3 Annual share of Global CO2 emissions from fossil fuels and Industry ……8
Figure 3. VAR & Multiple linear regression Framework illustrating workflow..... 46
Figure. 3.1 Head of Data pertaining to India – gdp, population and co2_including_luc ……. 47
Figure 3.2 Missing values in the dataset.……48
Figure. 4.1 Head of the dataset ……62
Figure. 4.2 Different columns in the dataset ……62
Figure. 4.3 Missing values (percentage) in the dataset ……62
Figure. 4.4 Heat map plot ……63
Figure. 5.1 Box plot of gdp ……...72
Figure. 5.2 Density plot of gdp …… 72
Figure. 5.3 Box plot of co2_including_luc ……...73
Figure. 5.4 Density plot of co2_including_luc ……73
Figure. 5.5 Variable values after applying Min - Max Scaler ……74
Figure. 5.6 OLS Regression results ……... ……75
Figure. 5.7 VIF results of variables ……... ……75
Figure. 5.8 Distribution of Error terms ……. ……...76
Figure. 5.9 Scatter plot of actual vs predicted values ……...77
Figure. 5.10 Regression plot of the model ……...77
Figure. 5.11 Difference values of Actual value and predicted value ……78
Figure. 5.12 Line plot of gdp, population and co2_including_luc ……79
Figure. 5.13 ADF test of population ……80
Figure. 5.14 ADF test of gdp ……...……81
Figure. 5.15 ADF test of co2_including_luc…...…….. 81
Figure. 5.16 Differenced ADF test of population ……81
Figure. 5.17 Differenced ADF test of gdp ……...……82

Figure. 5.18 Differenced ADF test of co2_including_luc…...…… 82
Figure. 5.19 Granger causality matrix …….83
Figure. 5.20 VAR Order Selection …….83
Figure. 5.21 OLS regression results of VAR ……...84
Figure. 5.22 Actual vs Forecast data of population …….85
Figure. 5.23 Actual vs Forecast data of gdp…….. …….85
Figure. 5.24 Actual vs Forecast data of co2_including_luc…….. …….86

LIST OF ABBREVIATIONS

ADF ……. Augmented Dickey-Fuller test
CCR…… Correct classification rate
COP…… Conference of Parties
co2_including_luc … CO2 emissions including land use change
GDP…… Gross Domestic Product
MLR ……. Multiple Linear Regression
LF-VAR……... Long Frequency Vector Auto Regression
luc…… Land use change
OLS…..... Ordinary Least Squares
RFE …… Recursive Feature Elimination
SDG…… Sustainable Development Goal
VAR……..... Vector Auto Regression
VIF ……. Variance Inflation Factor

TABLE OF CONTENTS

DEDICATION ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
LIST OF FIGURES v
LIST OF ABBREVIATIONS vi
CHAPTER 1: INTRODUCTION 5
1.1 Background of the Study 5
1.2 Related Research 10
1.3 Research Questions 14
1.4 Aim & Objectives 15
1.5 Significance of the Study 15
1.6 Scope of the Study 16
1.7 Structure of the Study 16
CHAPTER 2: LITERATURE REVIEW 18
2.1 Introduction 18
2.2 Introduction to CO2 Emissions 18
2.3 Relationship between GDP, Population and CO2 emissions 19
2.4 Stationary Test 37
2.4.1 Augmented Dickey - Fuller Test 37
2.5 Causality Test 38
2.5.1 Granger Causality Test 38
2.6 Introduction to Time Series Forecasting 39
2.6.1 Importance of Time Series Forecasting 40
2.6.2 Different types of Time Series Techniques 41
2.6.3 Different Types of Challenges and Considerations 42
2.7 Use cases of Time Series Forecasting in prediction of CO2 emissions 42
2.8 Summary 44
CHAPTER 3: RESEARCH METHODOLOGY 45
3.1 Introduction 45
3.2 Research Approach 45
3.3 Dataset Description 46

3.4 Pre-Processing and Exploratory Data Analysis 48
3.5 Multivariate Time Series Analysis 49
3.6 Vector Auto Regression (VAR) 49
3.6.1 Visualize and Analyse Time Series Data 51
3.6.2 Check for Stationarity 51
3.6.2.1 Unit Root Test 51
3.6.3 Granger's Causality Test 52
3.7 Split & Model the Data 53
3.8 Fit the data using the best lag value 54
3.9 Forecast, predict & plot the results 54
3.10 Evaluation 55
3.10.1 Root Mean Square Error 55
3.11 Multiple linear regression 55
3.12 Data Preparation 56
3.12.1 Handling Categorical Variables 56
3.13 Spilt the data into the train & test set 56
3.13.1 Feature Scaling 56
3.14 Building the Model 57
3.14.1 Variance Inflation Factor (VIF) 57
3.15 Residual analysis 58
3.16 Making Predictions 58
3.17 Model evaluation 58
3.18 Summary 59
CHAPTER 4: IMPLEMENTATION AND ANALYSIS 60
4.1 Introduction 60
4.2 Dataset Description 60
4.3. Dataset Preparation and Exploration 60
4.3.1 Identification of missing values 62
4.3.2 Univariate and Bivariate analysis 62
4.3.3 Treatment of missing values 63
4.4 Model Building 64
4.4.1 Feature Scaling 64

4.4.2 Recursive Feature Elimination 64
4.4.3 Variance Inflation Factor 65
4.4.4 Residual Analysis 65
4.4.5 Predictions and Evaluation 66
4.5 Vector Auto Regression 66
4.5.1 Stationary tests 67
4.5.2 Differencing 67
4.5.3 Causality 68
4.6 Building the model 68
4.6.1 Forecasting 68
4.6.2 Evaluation 69
4.7 Summary 69
CHAPTER 5: RESULTS AND DISCUSSION 71
5.1 Introduction 71
5.2 Dataset Issues 71
5.3 Univariate analysis results 71
5.4 Building the model – Multiple Linear Regression 74
5.4.1 Feature Selection 74
5.4.2 Recursive Feature Elimination & VIF results 74
5.4.3 Residual analysis 75
5.4.4 Model Evaluation 76
5.5 Vector Auto Regression 79
5.5.1 Exploratory data analysis 79
5.5.2 Stationary test results 80
5.5.3 Causality test results 82
5.5.4 Building the model 83
5.5.5 Forecasting 85
5.5.6 Evaluation 86
5.5.7 Comparison MLR vs VAR model evaluation 86
5.5.8 Summary 87
CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS 88
6.1 Introduction 88

6.2 Discussions and conclusions 88
6.3 Future Recommendations 90
References 91
APPENDIX A: RESEARCH PROPOSAL 95

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Climate change is mostly driven by the emissions of greenhouse gases (GHGs) that are produced by human activities. Approximately 60% of greenhouse gas (GHG) emissions are attributed to a mere 10 countries, whilst the 100 countries with the lowest emissions collectively account for less than 3% of total GHG emissions. Energy accounts for around 75% of world emissions, with agriculture being the subsequent major contributor. In the energy sector, the primary source of emissions is the electricity and heat generation sector, which is subsequently followed by the transportation and manufacturing sectors. The sector of land use, land use change and forestry (LULUCF) have a dual role as both a source and sink of emissions, making it a crucial component in achieving net-zero emissions. In the year 2020, the combined contributions of China and the United States accounted for 40% of the total GHG emissions on a worldwide scale. Subsequently, the European Union, India, the Russian Federation, and Indonesia followed suit in terms of their respective emissions (Washington, DC: World Resources Institute, 2022).

In the year 2020, China emerged as the primary source of carbon dioxide (CO2) emissions, accounting for around 30% of the global total. The United States followed closely behind, contributing approximately 12% of the global CO2 emissions, while India accounted for approximately 6.2% of the global emissions (Washington, DC: World Resources Institute, 2022). The Prime Minister of India, Narendra Modi, has expressed his dedication to attaining carbon neutrality by the year 2070. Concurrently, leaders from around the world have reached a collective consensus to restrict the increase in global average temperature to a level below 2°C. This commitment was established during the 26th Conference of the Parties (COP 26) convened in Glasgow.

The Earth's greenhouse gas emissions trap solar heat. This causes climate change and global warming. The earth is warming faster than ever. Warmer temperatures are altering weather patterns and nature's balance. This endangers us and all life on Earth.

Climate change increases heatwaves, strong storms, drought, and species extinction. Rising temperatures have caused more storms, deaths, and economic losses. Tropical cyclones,

hurricanes, and typhoons are fuelled by the warming ocean. Carbon dioxide absorption by the ocean threatens coral reefs and marine life. Climate change might wipe out one million terrestrial and marine species in the next few decades.

Fish, crops, and animals die or produce less due to climate change, causing world hunger. The ocean acidifies, reducing crops and livestock health. Climate change is humanity's biggest health risk, killing 13 million people annually and straining healthcare systems.

Floods destroy urban slums, make outdoor labour tougher, and reduce crop yields due to climate change. The most vulnerable and unprepared countries for climate change send the most refugees.

The study aims to examine the correlation between electricity consumption and real gross domestic product (GDP), with a specific emphasis on their impact on environmental deterioration, specifically in terms of carbon dioxide (CO2) emissions. The study employed annual time series data spanning from 1971 to 2017 to investigate the presence of causal linkages, both in the short and long term, using the Dickey-Fuller test, Johansen cointegration analysis, and Granger causality analysis. The research revealed a sustained correlation between energy consumption, economic growth, and carbon dioxide (CO2) emissions in the short term (Pandey and Rastogi,2019).

The impact of Indian CO2 emissions on economic growth, using the SDGs framework and 2030 targets. It examines GDP, energy intensity, and CO2 emissions, finding a one-way link between energy use and GDP. The research suggests that conservative energy measures may hinder economic growth due to energy dependence. The study suggests India should switch to renewable energy for cleaner energy and eco-friendly ecosystems, given global environmental awareness (Udemba et al., 2021).

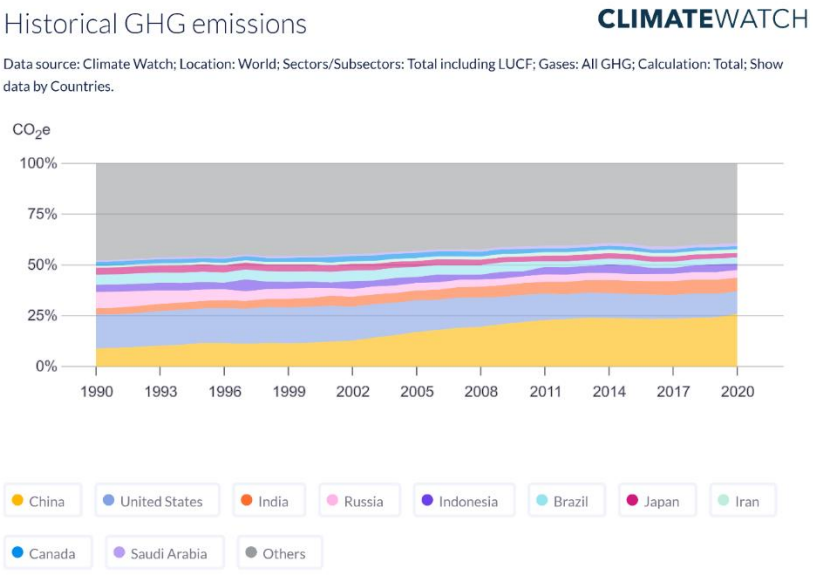


Figure. 1 GHG emissions country-wise percentage from 1990 – 2020 (Washington, DC: World Resources Institute, 2022)

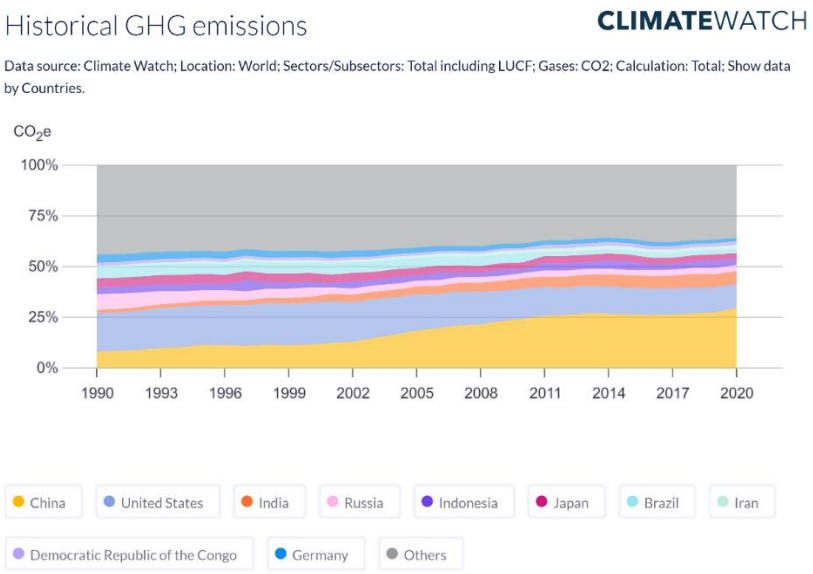


Figure. 1.1 CO2 emissions country-wise percentage from 1990 – 2020 (Washington, DC: World Resources Institute, 2022)

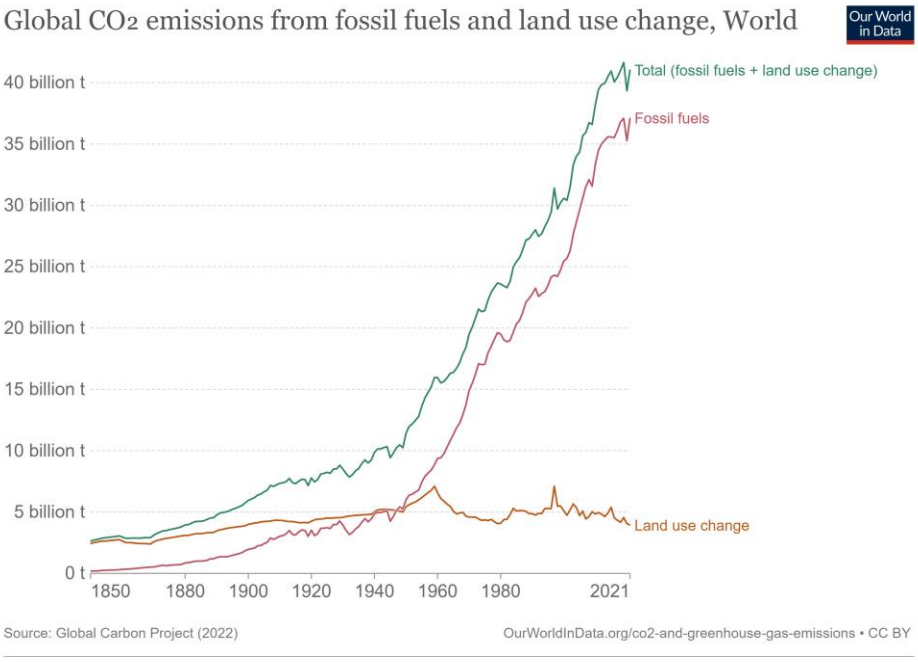


Figure. 1.2 Global CO2 emissions from fossil fuels and land use change (Ritchie, 2020)

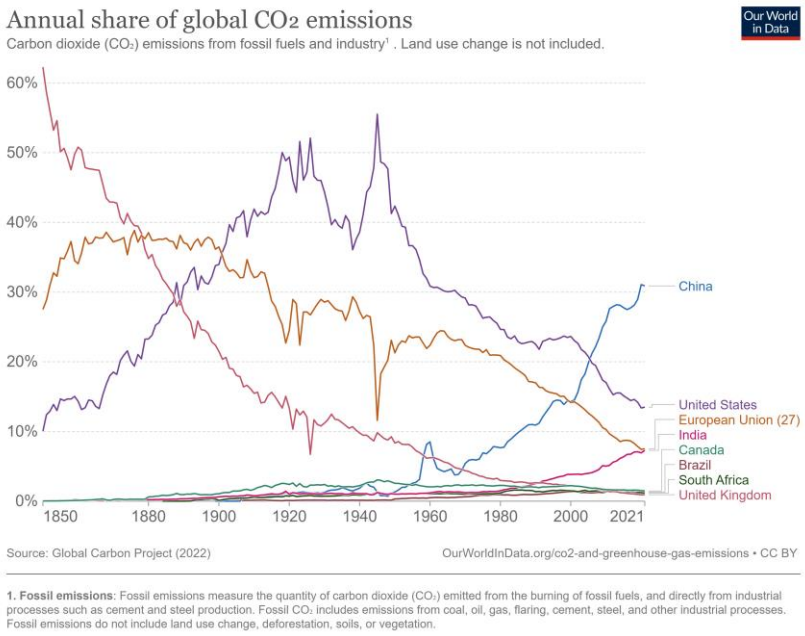


Figure. 1.3 Annual share of Global CO2 emissions from fossil fuels and Industry (Ritchie, 2020)

From Figure. 1.2, the observed trend entails the escalation of global carbon dioxide (CO2) emissions resulting from the utilisation of fossil fuels and alterations in land use, commencing from the mid-18th century and persisting to the present day. It is evident that there has been a notable increase in emissions stemming from fossil fuel usage, but emissions resulting from land use change have experienced a modest drop in recent years. In general, it can be observed that overall emissions have exhibited a state of relative stability during the previous decade.

Figure. 1.3 explains the temporal evolution of emissions has exhibited substantial alterations. The United Kingdom held the title of the world's largest emitter until 1888, at which point it was surpassed by the United States. The reason behind this phenomenon can be attributed to the United Kingdom's status as the pioneer in industrialization, a pivotal development that subsequently led to significant enhancements in the quality of life for a substantial portion of its populace.

Although the increase in carbon dioxide (CO2) emissions is associated with detrimental environmental effects, it is important to acknowledge that historically, these emissions have been a result of advancements in human living standards. However, it is imperative to acknowledge that the mitigation of CO2 emissions holds significant importance in safeguarding the quality of life for forthcoming generations. The consideration of both the environmental and human welfare effects of emissions is crucial in order to construct a future that is characterised by sustainability and the provision of elevated standards of life for all individuals.

The increase in emissions and improvement in living conditions in North America and Oceania occurred in close succession to advancements observed in the United Kingdom. Asia is currently home to a significant number of the world's top emitters. Nevertheless, the escalation of Asia's emissions has solely transpired inside the past several decades. This phenomenon may be attributed to significant advancements in living conditions. Over the course of the past seven decades, the average life expectancy in Asia has risen from 41 to 74 years. Moreover, there has been a substantial decline in the prevalence of extreme poverty within the region. Additionally, a noteworthy milestone has been achieved as the majority of the Asian population has had access to formal education for the very first time.

While it is imperative for all nations to collaborate, it is crucial that the primary contributors to emissions take decisive action. China, the United States of America, and the 28 member countries of the European Union collectively contribute to about 50% of the total world emissions. The world's ability to achieve its global targets will be significantly hindered in the absence of commitment from these major emitters.

1.2 Related Research

The association between electricity use and real GDP is examined in this empirical study. This relationship's effects on environmental degradation, particularly CO2 emissions, are also examined. This study examines the links between real GDP, electricity usage, and CO2 emissions. The investigation used 1971–2017 annual time series data. The Dickey-Fuller test determines variable stationarity. Johansen cointegration and Granger causality are used to examine the short- and long-term causal relationships between power consumption, real GDP, and CO2 emissions. Cointegration among specific combinations of the two variables is shown by the Johansen cointegration test, demonstrating a long-term link between the defined variables. Energy use also causes economic development and CO2 emissions in the near run. (Pandey and Rastogi,2019).

Long-term relationships between health expenditure, CO2 emissions, and GDP per capita in 18 OECD countries were examined using annual time-series data from 1975 to 2017 using McNown et al.'s bootstrap autoregressive-distributed lag (ARDL) cointegration model. The Netherlands, New Zealand, and the US have cointegration in real GDP per capita, health expenditure, and CO2 emissions. The major results reveal a short-run relationship between the three variables. Health spending and GDP growth in Germany and the US, CO2 emissions and GDP growth in Canada, Germany, and the US, and health expenditure and CO2 emissions in New Zealand and Norway are bidirectionally correlated. The results reveal unidirectional causality in other nations (Wang et al., 2019).

Carbon dioxide emissions were examined in relation to energy use and economic growth. Regression, pooled OLS regression, fixed effects, Granger causality, and panel cointegration tests were used in the study. Data from 70 nations from 1994 to 2013 was analysed. The study

found a reciprocal causal relationship between the population, capital stock, economic growth, and CO2 emissions, but not energy consumption. Cointegration tests show economic growth, CO2 emissions and energy use are linked. The study stressed the need for a global transition to mitigate carbon emissions and prioritise climate financing, including the mobilisation of public, private, and alternative financial resources to invest in renewable energy and environmentally sustainable projects. (Osobajo et al., 2020).

Environmental deterioration in developing countries is caused by using non-renewable energy for economic growth. This study uses 1965–2015 yearly time series data to examine Pakistan's economic growth, energy consumption, and CO2 emissions. The ARDL estimations show that economic growth and energy consumption increase CO2 emissions in Pakistan, both short-term and long-term. These findings suggest that Pakistani governments should encourage and utilise renewable energy sources to address rising energy demand. Replace coal, gas, and oil with renewable energy to cut CO2 emissions and sustain Pakistan's economy (Khan et al., 2020).

Indian CO2 emissions, a major polluter, harm the ecology despite the expansion. Indian and other CO2 emissions and economic growth literature is divided. The SDGs framework and 2030 targets are used to study India's energy, climate, and economic growth. The study investigates Indian GDP, energy intensity, and CO2 emissions. Trade openness and energy use reduce omitted variable bias in a carbon income function. The Autoregressive Distributed Lag (ARDL) method and modified Wald test of Toda Yamamoto (T-Y) are used for 1975–2017 yearly time series data. Long-term equilibrium exists for the variables. CO2 emissions statistically damage trade and growth. Energy-induced economic expansion supports ARDL regression. Energy use and GDP have a one-way link, suggesting that conservative energy measures will hurt economic growth owing to energy dependence. Given global environmental awareness, India should switch to renewable energy for cleaner energy and eco-friendly ecosystems (Udemba et al., 2021).

State-level CO2 emissions, energy consumption, and GDP from 1997 to 2016 are examined. GDP and energy consumption—total, non-renewable, renewable, industrial, and residential—affect CO2 emissions among states, according to static and dynamic models. Both static and dynamic models reveal a long-term relationship between state-level CO2 emissions and energy use. Non-renewable, industrial, and residential energy use raises CO2 emissions, while renewable energy decreases them. CO2 emissions and GDP have an inverted-U relationship,

supporting the Environmental Kuznets Curve (EKC) hypothesis across states. These results are consistent and robust across states using static and dynamic models. This can assist policymakers reduce CO2 emissions in all U.S. states. (Salari et al., 2021).

Urbanisation, energy use, gross capital formation, CO2 emissions, and economic growth in South Korea are examined considering the UN Sustainable Development Goals (SDGs), which emphasise energy access (SDG-7) and sustainable development (SDG-8). This link has not been studied using advanced econometrics. The study uses 1965–2019 data. This association is studied using ARDL, DOLS, and completely modified OLS. Gradual shift and wavelet coherence dictate causation. The ARDL bounds test links variables of interest long-term. CO2 emissions significantly harm economic growth, emphasising the need for South Korea to switch to renewable energy to promote sustainable energy and a sustainable ecosystem. Conservative energy policies may restrict economic growth, supporting the energy-induced growth hypothesis. A one-way causation between energy use and GDP suggests South Korea should avoid stringent energy laws that could limit economic growth. This significantly impacts South Korean GDP and macroeconomic indices (Adebayo et al., 2021).

The growth of Vietnam's per capita income from 1990 to 2019 was evaluated in relation to the use of non-renewable energy, renewable energy, and CO2 emissions. The study cointegrated yearly Vietnam data using Autoregressive Distributed Lag (ARDL). Long-term, non-renewable energy use raises per capita income, whereas CO2 emissions lower it. Non-renewable and renewable energy use changes enhance Vietnam's per capita income. Prior non-renewable energy usage changes hampered Vietnamese income growth. This study illuminates the growth effects of renewable, non-renewable, and CO2 emissions. The findings help Vietnam establish a long-term economic growth strategy. (Nguyen and Le, 2022).

Economic growth, industrial production, and energy consumption determine per capita carbon dioxide emissions, which the proposed study evaluates. Countries with high worldwide CO2 emissions per person are the focus of this study. Data were analysed using panel regression with heterogeneous time trends. After extensive examination, a non-effect panel regression model with heterogeneous time trends is best for our study. (Puntoon et al.,2022) found a positive link between energy use and CO2 emissions. However, economic development and industrial production have a weaker link with CO2 emissions. This study examines how economic

development, industrial production, and energy consumption affect CO2 emissions per capita. The study analyses nations with the highest per-capita CO2 emissions using panel regression with heterogeneous time trends. After careful consideration, a non-effect panel regression with heterogeneous time trends is best. Energy consumption has a strong positive effect on CO2 emissions, while economic expansion and industrial production have a weaker effect.

India's economic growth, energy consumption, FDI, CO2 emissions, population density, inflation, and agricultural land are assessed. From 1985 to 2019, autoregressive distributed lag models use annual time-series data. Fully modified ordinary least squares, dynamic OLS, canonical co-integrating regression, variance decomposition, and impulse response function show the model's resilience and uniqueness. Variable cointegration implies a long-term relationship. Energy use, CO2 emissions, inflation, and farmland effect short-term growth. Inflation and agricultural land hurt long-term economic growth, but carbon emissions help. Granger causality links economic growth to energy, carbon, and agriculture. The research enhances policymakers' economic growth indicator understanding and literature. Policymakers should set renewable energy consumption goals to boost growth and reduce carbon emissions. Eco-friendly and economically sustainable futures are supported. (Singh and Kaur, 2022).

China's 1990–2020 renewable energy consumption, output, export, and CO2 emissions are examined. Econometrics supports feedback theory with a two-way causal link. In the medium term, industrial and agricultural export and renewable energy consumption are negatively correlated, supporting growth theory. This means peak export demand may require additional fossil fuels due to renewable energy supply constraints. The findings back China's long-term pollution control and renewable energy objectives. (Hao, 2022).

The relationship between GDP CO2 intensity and environmental deterioration in the rising nation Turkey from 1990 to 2018 is examined. Economic development, foreign direct investment, and renewable energy usage are controlled to investigate their consequences. The bounds test fully modified ordinary least squares (FMOLS), Gregory and Hansen cointegration test, dynamic OLS, nonlinear autoregressive distributed lag (NARDL) model, and CCR are used to evaluate these factors' effects on environmental degradation in Turkey. Environmental degradation in Turkey is linked to GDP CO2 intensity, and lowering GDP intensity lessens it, according to empirical studies. Economic growth also promotes environmental sustainability.

These findings suggest that aggressive government measures can tackle environmental issues (Abbasi et al., 2022).

In Taiwan, the MF-VAR model is being applied for the first time. Using a high-frequency dataset, economic development and carbon dioxide emissions are investigated from 1970 to 2019. Energy use is primarily employed as a control variable to mitigate the influence of confounding variables that may be absent from the study. Based on forecast error variance decomposition and the Granger causality test, the empirical results suggest that the MF-VAR model shows variable relationships better than the classic VAR model. The empirical LF-VAR model demonstrates a positive relationship between primary energy consumption and economic growth, leading to subsequent increases in CO2 emissions. The MF-VAR model in Taiwan reveals a bidirectional and causal association among economic progress, carbon dioxide emissions, and primary energy use. The findings of this analysis indicate that the implementation of a robust energy plan is important for Taiwan to enhance its economic growth (Chang et al., 2023).

G7 CO2 emissions and economic growth are examined from 1820 to 2021. Emissions and economic growth asymmetry are examined using quantile-vector autoregression at different distribution points. GDP increase asymmetrically affects CO2 emissions. This implies that extreme quantiles and the median affect CO2 emissions and economic growth. CO2 emissions and economic growth are bidirectional, says the study. Quantile asymmetry and variance in G7 nations' transmission of effects between variables are also shown. Between 1850 and 2000, CO2 emissions were a net transmitter throughout North America, encompassing the US and Canada. In contrast, (Jebabli et al.,2023) observed that CO2 emissions in Europe (UK, Italy, France, and Germany) and Japan initially acted as net receivers and then as net transmitters.

1.3 Research Questions

The research questions corresponding to each of the mentioned study objectives are as follows:

1. How carbon dioxide (CO2) emissions, population size, and economic growth are correlated?

2. Which of the factors, population, or economic growth, exhibits a causal relationship with CO2 emissions?
3. Which model MLR or VAR gives highest performance of accuracy in terms of evaluation?

1.4 Aim & Objectives

The main objective of this study is to examine the relationship between population growth, economic development, and CO2 emissions from 1850 to 2021. This study aims to investigate the link between economic development and environmental deterioration.

The research objectives are formulated based on the aim of this study which is as follows:

- To analyze and identify the relationship between economic growth and environmental degradation.
- To explore the causality between economic growth and environmental degradation.
- To assess the VAR & Multiple Linear regression models and identify the most precise one to find the relationship between economic growth and environmental degradation.
- To use the Vector Autoregressive Model of Timeseries Modeling to forecast CO2 levels of actual values vs predicted values from 2011 to 2021 based on historical data.

1.5 Significance of the Study

The primary objective of this study is to examine the correlation between carbon dioxide (CO2) emissions and economic growth through the utilisation of Vector Autoregression (VAR) and Multiple Linear Regression models. To ascertain the primary determinant of carbon dioxide (CO2) emissions in India, it is imperative to investigate the relative influence of two key factors: gross domestic product (GDP) and population size. The study's findings have important policy consequences for the government. Indian Prime Minister Modi has committed to achieving carbon neutrality by the year 2070, while global leaders have collectively agreed to limit the rise in global average temperature to below 2°C. This commitment was made during the COP 26 meeting held in Glasgow. The primary cause of climate change and the rise in global temperatures can be attributed to the significant contribution of greenhouse gas emissions. Carbon dioxide (CO2) emissions make a significant contribution to this phenomenon. To

determine the best appropriate machine learning algorithm for our specific use case, we will do a comparative analysis and review relevant literature references.

1.6 Scope of the Study

World in Data dataset on CO2 emissions is used to select the best model to determine the relationship between CO2 emissions and economic growth. The research will make use of data on India's GDP, population, and CO2 emissions including changes in land use from 1850 to 2021. Carbon dioxide (CO2) emissions, which encompass land use change (luc), primarily consider territorial emissions, while neglecting emissions associated with traded goods. Land use change refers to the process by which humans modify the purpose of a given area of land, resulting in its conversion from one use to another. The study's scope is restricted to comparing the VAR model with the Multiple linear regression model, to identify the most accurate model for the study through model evaluation.

1.7 Structure of the Study

For ease of understanding the interim report has been divided into three chapters.

- **Introduction:** This chapter analyses the subject matter and explains why it was chosen as the study's main emphasis. This chapter addresses the problem statement and reviews relevant literature to identify gaps and restrictions. The chapter shows the extent and importance of the current study and its ability to provide field insights by identifying these inconsistencies. This chapter also describes the study's setting, identifying its bounds and explaining its importance in correcting earlier research inadequacies. It can also define the study's goals and research questions, making it clearer and more focused.
- **Literature review:** Starting with the chapter's agenda is vital. This chapter provides an introduction to greenhouse gases, focusing on CO2 emissions. This chapter discusses CO2 emissions' goals. Burning fossil fuels is the main source of residential and commercial CO2 emissions. It also examines GDP, population, and CO2 emissions using various modelling methodologies and their constraints. Current research gaps can be identified by this analysis. The sub-chapter discusses Stationary data and Causality tests

for GDP, Population, and CO2 emissions Time Series data. It covers Time Series Forecasting's importance, understanding Time Series data and forecasting, and Time Series data analysis methodologies. Different challenges and concerns are crucial while analysing Time Series data. Time Series Forecasting usage in CO2 emission prediction and how it helps governments and organisations make policy changes and prevent climate change.

- **Research Methodology:** The study focuses on constructing a resilient model using historical data on CO2 and greenhouse gas emissions in India. It outlines the research approach, dataset selection, pre-processing and cleaning strategies, algorithm selection for time series prediction, model architectural design, and performance evaluation metrics for precision and efficacy. The chapter also provides an overview of the selected dataset and its corresponding procedures.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

In the initial stage, it is crucial to outline the agenda of this chapter. This chapter serves as an introduction to the Greenhouse gases specifically focussing on CO2 emissions. This chapter covers various purposes by which CO2 emissions happen. The primary source of CO2 emission from both residential and commercial properties is burning of fossil fuels. Furthermore, it explores the relationship between GDP, Population and CO2 emissions in terms of various modelling methods and highlighting their limitations. This analysis allows for identifying gaps in the existing research conducted so far.

The sub-chapter explores the understanding of Stationary data and Causality tests in terms of GDP, Population and CO2 emissions Time Series data. It talks about the importance of Time Series Forecasting, Understanding of Time Series data and Forecasting, Different types of methods and techniques used for analysis of data using Time Series. Different types of challenges and considerations are of prime importance and needs to be addressed when Time Series data analyses needs to be carried out. Various use cases of Time Series Forecasting in prediction of CO2 emissions and how they play a vital role in helping Governments, Organisations to make policy changes and mitigate climate change.

2.2 Introduction to CO2 Emissions

A greenhouse gas (GHG) refers to a gaseous molecule that possesses the ability to absorb and emit infrared radiation. This characteristic results in a reduced amount of heat escaping into space, effectively confining it within the lower atmosphere. The primary greenhouse gases present in the Earth's atmosphere consist of water vapour, carbon dioxide (CO2), methane (CH4), nitrous oxide (N2O), and ozone (O3).

The fundamental catalyst for global climate change is the release of carbon dioxide. The imperative to mitigate the most severe consequences of climate change necessitates the expeditious reduction of global emissions, a widely acknowledged proposition.

Carbon dioxide (CO2) emissions arising from the combustion of fossil fuels for the subsequent purposes:

- The power industry primarily relies on the combustion of fossil fuels, particularly coal and natural gas, for the generation of electricity. The primary source of greenhouse gas emissions from industrial activities stems from the combustion of fossil fuels for energy generation.
- Additional Industrial Processes
- Transportation mostly relies on the combustion of fossil fuels to power various modes of conveyance, including automobiles, trucks, ships, trains, and aircraft.
- Non-combustion refers to a range of chemical reactions that are essential for the manufacturing of goods using raw materials. Examples of such reactions include cement production, the utilisation of limestone and dolomite for carbonation purposes, the non-energy utilisation of fuels and other combustion processes, chemical and metal processes, the use of solvents, agricultural practices involving liming and urea, as well as the occurrence of fires involving waste and fossil fuels.

The structures under question encompass both commercial and residential properties. The primary sources of greenhouse gas emissions in both commercial and residential sectors stem from the combustion of fossil fuels for heating purposes, the utilisation of certain items that include greenhouse gases, and the management of waste materials.

The concentration of carbon dioxide (CO2) in the Earth's atmosphere has experienced a notable increase because of human activity, particularly since the onset of the Industrial Revolution.

2.3 Relationship between GDP, Population and CO2 emissions

While there is a substantial body of literature that has examined the impact of technology advancements on energy consumption and carbon dioxide (CO2) emissions, there has been less research focused on understanding the variations across different sectors in terms of the relationship between technological progress and CO2 emissions. This study aims to address the existing gaps in the current body of research by conducting an empirical analysis to examine the

influence of technical advancements on carbon dioxide (CO2) emissions. Furthermore, it seeks to consider the diverse impact of these advancements across different economic sectors. This study initially examines the impact of technical advancements within various sectors, including heavy industry, light industry, construction industry, and service industry. The study used a panel quantile regression methodology and a balanced city panel data model to analyse data from China spanning the years 2001 to 2013. The empirical findings indicate that the advancement of technology in both heavy industry and light industry unexpectedly leads to increased levels of CO2 emissions, despite its notable role in enhancing energy efficiency. In contrast to its impact on heavy and light industries, technical advancements in the construction and service sectors have a detrimental influence on carbon dioxide (CO2) emissions. Furthermore, this study also reveals that the magnitude of the various effects of technology advancement is contingent upon the degree of emitters. This research aims to provide a thorough knowledge of the relationship between CO2 emissions and technical progress by examining the varied effect that technological progress has on different economic sectors and emitters (Wang et al., 2019).

This study examines the correlation between energy consumption, economic growth, and carbon dioxide (CO2) emissions within the framework of the Environmental Kuznets Curve. This study examines the oil-rich countries in the Middle East and North Africa (MENA) region, including Algeria, Bahrain, Iran, Kuwait, Oman, Qatar, and Saudi Arabia, from 1995 to 2014. CO2 emission is expected to follow a quadratic model formulation in relation to both energy consumption and economic growth. The utilisation of multivariate linear regression is employed to evaluate the confirmation of the Environmental Kuznets Curve (EKC). The objective of this study is to examine the relationship between energy consumption and carbon dioxide (CO2) emissions, GDP squared (GDP2), and gross domestic product (GDP). The problem at hand can be examined using either a panel data framework or a multivariate regression framework. The former method is extensively utilised in the fields of econometrics and social sciences, whereas the later method, despite its theoretical equivalence, flexibility, and user-friendly nature, has not garnered sufficient attention. This work aims to enhance the understanding of the existence of other techniques for analysing identical challenges. The Environmental Kuznets Curve (EKC) can be seen as a theoretical proposition, but failure to adequately verify the underlying statistical assumptions can result in incongruous findings. This article offers a systematic approach to effectively analyse the problem within a multivariate regression framework and assess the

validity of its assumptions. The MATLAB instructions that are pertinent to the topic are also provided within the publications (Ardakani and Seyedaliakbar, 2019).

Iran has become a major CO2 emitter in recent decades. The nation emits less carbon dioxide than Japan and Germany. Iran has a lower GDP than Berlin and Tokyo combined. Additionally, crude oil sales account up a large portion of Iran's revenue. Thus, Iran's high energy intensity explains its significant CO2 emissions. In the meanwhile, the administration has no coherent plan. The Sixth Five-year Development Plan of Iran contains aggressive energy intensity, GDP growth, and renewable energy goals. Notably, CO2 emissions are not addressed. Thus, a quick conflict resolution seems unlikely. This analysis predicts Iran's 2030 CO2 emissions. BAU and the Sixth Development Plan will be considered to achieve this. MPR and MLR will be used in the analysis. Based on the business-as-usual scenario, Iran is unlikely to meet its Paris Agreement responsibilities. If the strategically intended sustainable development plan were fully implemented, it may have reached the aim by 2018. (Hosseini et al., 2019).

This research endeavour aims to elucidate the interplay between per capita GDP, CO2 emissions, and energy consumption in Vietnam, employing an econometric methodology. The analysis involved the utilisation of annual data spanning from 1970 to 2014. Various statistical techniques were employed, including tests for stationarity, identification of structural breaks, the Toda-Yamamoto test, the Johansen and Juselius approach, and variance decomposition. The findings of our analysis indicate that there is a unidirectional causal relationship, with economic expansion exerting an influence on energy use. The significance of this outcome lies in its support for the conservation hypothesis pertaining to the economy of Vietnam. The findings from the variance decompositions provide evidence to refute the notion that energy is neutral for growth. Instead, they suggest the presence of a relationship between energy and growth, but with effects that may be temporary in nature (Morelli and Mele, 2020).

This study uses a data-driven nonparametric additive regression model to evaluate how fossil energy abundance affects China's CO2 emissions and economic growth. An inverted "U-shaped" relationship exists between eastern fossil energy abundance and economic expansion. Investments in oil processing, coal mining, and coking fluctuate. In contrast, rich fossil fuels have a positive nonlinear "U-shaped" influence on core region economic growth. This shows that

fossil fuels did not boost economic growth in the early stages. Later, it had a greater impact. The abundance of fossil energy has a "U-shaped" positive effect on CO2 emissions in the east and centre. Coal and oil consumption fluctuated over time. In the west, fossil energy consumption has an inverted "U-shaped" effect on CO2 emissions. This is because natural gas and oil use and production are phased differently. (Lin and Xu, 2020).

Worldwide, carbon dioxide (CO2) emissions, a greenhouse gas, are the main cause of global warming and environmental degradation. This study focuses on China, the world's largest carbon emitter. This study examines Chinese economic growth and CO2 emissions from 1950 to 2016. Wavelet coherence will be used to examine long-term and short-term causal relationships between variables. This study uses wavelet coherence, Maki cointegration, Fourier Toda-Yamamoto causality, and nonparametric Granger causality tests to examine the long-term causal relationship between CO2 emissions and economic growth. This study found a medium-term and short-term link between CO2 emissions and economic growth in the 2000s. China has a long-term cointegration between economic growth and CO2 emissions. The study also shows that Chinese economic growth predicts CO2 emissions in the short and medium run. In the 1980s and 1990s, economic growth was positively correlated, but only temporarily. The Fourier Toda-Yamamoto causality, Toda-Yamamoto causality, and nonparametric Granger causality tests show that economic growth is a credible policy signal for China's CO2 emissions. (Kirikkaleli, 2020).

The interconnection between energy consumption, population dynamics, carbon emissions, and the implementation of carbon taxes should not be disregarded. This link holds significant importance for environmentalists, economists, policy makers, and scholars. This study aims to examine the interrelationship between carbon emissions, population, energy consumption, and carbon taxation throughout the period of 1970 to 2018. The primary aim of this study was to investigate the relationship between carbon emissions, carbon taxation, energy consumption, and population. The objective was successfully accomplished by employing the autoregressive distributive lag model (ARDL), which is known for its ability to provide precise parameter estimates. The findings of the study indicate that there is a positive correlation between energy consumption and population increase and carbon dioxide emissions. Conversely, the implementation of a carbon tax has been observed to have a mitigating effect on carbon

emissions. According to the findings of the study, it is recommended that the government of South Africa take measures to encourage the adoption of clean energy and devise strategies to mitigate population increase as a means of reducing carbon emissions (Garidzirai, 2020).

This research examines the effects of carbon dioxide (CO2) emissions, population density, and trade openness on the economic growth of five nations in South Asia. The panel co-integration approach is employed in this study, utilising data spanning from 1990 to 2017. The focus of the analysis is on the extended neoclassical growth model. The results produced from the analysis indicate that there is a positive relationship between CO2 emissions and population density, whereas there is a negative relationship between trade openness and economic growth in the South Asian region. The impact of population density surpasses that of CO2 emissions in terms of magnitude. The findings of the Granger causality analysis demonstrate the presence of a reciprocal causal relationship between economic growth and CO2 emissions, as well as between trade openness and CO2 emissions. There exists a unidirectional relationship whereby trade openness influences economic growth, population density affects CO2 emissions, and both labour and economic growth impact population density. The findings are utilised to produce a comprehensive policy recommendation (Rahman et al., 2020).

The objective of this research is to examine the nonlinear relationship between urban population, energy consumption, economic growth, and carbon emissions in specific African economies throughout the time frame of 2005 to 2019. The inquiry employs the non-linear panel smooth transition regression (PSTR) estimate technique. The findings indicate a rejection of the null hypothesis positing linearity, in favour of the presence of non-linearity. The findings offer empirical support for the existence of the environmental Kuznets curve (EKC). Energy use has a direct correlation with carbon emissions in both systems. The rise in urban population is associated with a decrease in carbon emissions. The study suggests that governments should adopt policies focused on attaining low carbon mechanisms, such as green infrastructure and renewable energy systems, in order to expedite economic growth. These mechanisms have the potential to decrease energy consumption and mitigate greenhouse gas emissions. Hence, it is of utmost significance for the chosen African economies (Mosikari and Eita, 2020).

The main objective is to examine the correlations between energy consumption and carbon dioxide emissions and economic growth and CO2 emissions from 1970 to 2017. The second goal of this study is to determine the long-term relationship and causality between variables. The Granger causality test was used to achieve this. The data suggest urban population and energy demand are reciprocally linked. The third goal of this study is to examine how urbanisation, energy consumption, and economic expansion affect Nigerian carbon dioxide emissions. The study used autoregressive distributed lag (ARDL) testing to achieve this. Over a short period of time, energy consumption, the lagged effect of economic growth, and carbon dioxide emissions in Nigeria are positively and statistically significant. Urban population is the only element that negatively impacts Nigeria's CO2 emissions. The statistically substantial, negative link between urbanisation and CO2 emissions persists. Alternatively, energy consumption and economic expansion positively impact CO2 emissions. This issue is caused by the country's heavy use of non-renewable energy. The study argues that appropriate measures and mitigation methods are needed to reduce environmental harm and prevent further destruction. (Akorede and Afroz, 2020).

For an economy to achieve significant growth, it is often necessary to navigate a trade-off between the advancement of financial development and the potential deterioration of the environment. In the context of a nation such as Singapore, which has experienced significant economic growth and is renowned for its high population density, it is imperative to examine the impact of innovative green technologies on the attainment of economic excellence while minimising environmental costs. Using the innovative bootstrap autoregressive-distributed lag (BARDL) technique and analysing a time series dataset spanning from 1990 to 2018, the findings indicate a statistically significant positive association between green technology innovation and economic growth, as well as a statistically significant negative association between green technology innovation and carbon emissions. These relationships hold true in both the long run and short run. The outcomes of the study were analysed in order to derive many management consequences. Additionally, the limits of the study were considered, leading to the identification of potential possibilities for future researchers (Meirun et al., 2020).

Many governments worry about the greenhouse effect's economic and health effects. Transport contributes significantly to global GHG emissions. The transport sector emits 15% of global

greenhouse gas (GHG) emissions and 20% of energy-related CO2. Quantifying greenhouse gas (GHG) emissions from the road transport sector helps evaluate car energy consumption and suggest technological solutions to improve vehicle efficiency and reduce energy supply-related greenhouse gas emissions. This study aims to anticipate road transport industry greenhouse gas (GHG) emissions using a conceptual framework. Vehicle-Kilometre by Mode (VKM) and Number of Transportation Vehicles (NTV) are compared for six modes. This study examines motorcycles, passenger cars, tractors, single-unit trucks, buses, and light trucks. This investigation used 22-year data from the North American Transportation Statistics (NATS) web database. Multivariate regression and double exponential methods are used to model greenhouse gas (GHG) emissions. The findings show that VKM to NTV across transport modes significantly affects greenhouse gas (GHG) emissions. Adjusted R2 and R2 values measure the model's variance explanation at 89.46% and 91.8%, respectively. This suggests that VKM and NTV are the main causes of greenhouse gas (GHG) emissions. The model is used to assess different future plug-in hybrid and battery electric car scenarios. The deployment of battery electric vehicles would reduce CO2 emissions by 62.2%. This study will help develop effective road transport industry greenhouse gas (GHG) emission mitigation plans, regulations, and tactics. (Alhindawi et al., 2020).

The objective of this research is to examine the correlation between carbon dioxide (CO2) emissions, energy consumption, and economic growth (Gross Domestic Product, GDP) in the United States at the state level from 1997 to 2016. This research employs a range of quantitative methodologies, encompassing both static and dynamic models, to assess the effects of GDP and various forms of energy consumption (including total, non-renewable, renewable, industrial, and domestic energy) on carbon dioxide (CO2) emissions at the state level. The findings indicate the presence of a sustained association between different forms of energy consumption and CO2 emissions at the state level, as demonstrated by both static and dynamic models. The relationship between CO2 emissions and energy consumption can be characterised by a positive impact for total, non-renewable, industrial, and residential energy consumption, however a negative association is observed for renewable energy consumption. The results indicate a curvilinear association between carbon dioxide (CO2) emissions and gross domestic product (GDP), supporting the validity of the Environmental Kuznets Curve (EKC) theory when applied to different states. The findings exhibit strong consistency across different states when employing

both static and dynamic models. The findings of this study can be utilised by policymakers to establish relevant policies aimed at mitigating CO2 emissions within the various states of the United States (Salari et al., 2021).

The present research examines the relationship between carbon dioxide emissions (CO2E), gross domestic product (GDP), energy consumption (ENU), and population growth (PG) in India from 1980 to 2018 by employing the vector error correction model (VECM) and the autoregressive distributed lag (ARDL) approach. The unit root test, Johansen multivariate cointegration, and Variance decomposition analysis utilising the Cholesky technique were employed in our study. The Vector Error Correction Model (VECM) and Autoregressive Distributed Lag (ARDL) bound testing methodologies for cointegration propose the existence of a long-term equilibrium relationship among GDP, energy consumption, population increase, and CO2 emissions. The empirical results demonstrate the presence of a sustained equilibrium relationship among the variables. The findings of the Granger causality analysis indicate the presence of short-term bidirectional causation between GDP and ENU. Additionally, a unidirectional causality is observed between CO2E and GDP, CO2E and ENU, CO2E and PG, as well as PG and ENU. The variance decomposition analysis reveals that a significant portion, specifically 58.4%, of the forthcoming fluctuations in CO2E can be attributed to variations in ENU. Additionally, changes in GDP account for a modest 2.8% of the future fluctuations, while changes in PG contribute a mere 0.43% to these fluctuations. The ARDL test results demonstrate that a 1% rise in PG is associated with a 1.4% increase in CO2E. The present study examines several significant policy consequences (Pachiyappan et al., 2021).

This research examines the interrelationship between industrialization, economic growth, and carbon dioxide (CO2) emissions in the Chinese economy, while also considering the factors of trade openness and population density. Additionally, it assesses the concept of the environmental Kuznets curve (EKC). The findings from the calculations indicate that population density, industrial activity, and commerce have a positive impact on the levels of CO2 emissions in China. Conversely, the per capita GDP has a negative effect on CO2 emissions over a sustained period of time. The study also discovered a reciprocal causal association between CO2 emissions and industrialization, as well as a one-way relationship between population density and trade openness structure. The process of variance decomposition involves analysing the impact of

many factors on Chinese CO2 emissions, specifically focusing on the time-lag effect of CO2 emissions, industrialization, and per capita GDP. These factors are considered significant predictors of Chinese CO2 emissions. The current study is expected to initiate a discourse within the academic community, as it delves into the topic of the short-term and long-term consequences for China. Additionally, it offers policy recommendations for policymakers (Aslam et al., 2021).

Human-caused CO2 emissions are the main cause of global warming. This study intends to thoroughly examine human variables' impact on CO2 emissions. This study examines how population ageing, life expectancy, density, unemployment, GDP, and urbanisation affect per capita CO2 emissions. Data from 154 countries is analysed using linear panel data analysis and panel threshold regression. The link between unemployment and per capita CO2 emissions and urbanisation and CO2 emissions has no threshold value. This suggests a linear relationship between unemployment and urbanisation and per capita CO2 emissions. Urbanisation increases per capita CO2 emissions, but unemployment rates offset this. Additionally, there are threshold values that define the relationship between ageing, life expectancy, population density, and GDP and per capita CO2 emissions. This suggests a nonlinear link between per capita CO2 emissions, ageing, life expectancy, population density, and GDP. GDP has a decreasing effect on CO2 emissions per capita. Population ageing and higher life expectancy constrain per capita CO2 emissions, with the magnitude increasing with ageing and life expectancy. Increased population density reduces per capita CO2. Empirical study on 154 nations supports the idea that human factors affect carbon emissions in complex ways. (Wang and Li, 2021).

In recent years, Shandong Province has emerged as a significant contributor to China's carbon emissions. Nevertheless, previous research has not adequately examined the current patterns and primary determinants of this phenomenon at the municipal level. This study utilised accounting methods and the Logarithmic Mean Divisia Index (LMDI) to calculate the CO2 emissions at the city level. The objective was to obtain a comprehensive understanding of the contributing variables to CO2 emissions across 16 cities in Shandong Province from 2010 to 2018. The findings of the research demonstrate that there was a consistent upward trajectory in the levels of carbon dioxide (CO2) emissions in the region of Shandong from 2010 to 2018, with the exception of the year 2013. The GDP per capita and population size of Shandong Province have

contributed to the increase in energy-related CO2 emissions between 2010 and 2018. The primary factor contributing to the substantial increase in CO2 emissions in Shandong is energy intensity, which is closely followed by the energy consumption structure. The increase in CO2 emissions is partially mitigated by the emission intensity and geographical structure. The industrial structure has a crucial role in the reduction of emissions, although its impact on emission reduction is not consistently stable across various cities and sectors. This is particularly evident in the nonmetal and metal industry, petroleum and chemical industry, and energy sector. From 2010 to 2018, Dongying has consistently ranked as the highest emitter among all regions in Shandong province. The primary sources of emissions for this substance are predominantly derived from the petroleum and chemical industry. The primary determinants of influence are the energy intensity and industrial structure. The examination of carbon dioxide (CO2) emissions at the municipal level leads to a compelling suggestion that the various regions within Shandong Province should collaborate in order to enhance their growth patterns (Yang et al., 2021).

Air pollution can seriously harm health. Economic activity and non-renewable resource use can pollute the environment. CO2 emissions have been used to quantify environmental harm in previous research. CO2 emissions are rising in many countries, both developed and developing. In response to environmental concerns, the study examines how energy consumption, economic growth, and population growth in rural areas affect carbon dioxide emissions. This study aims to determine how rural population growth affects CO2 emissions. The panel autoregressive distributed lag (ARDL) method is used to analyse 1990–2015 data from nine rising nations in different areas. The data show that energy use and economic growth increase CO2 emissions over time. In contrast, rural population expansion does not affect CO2 emissions. Rural population growth does not affect CO2 emissions in the medium run. However, energy use and economic growth might harm the environment in the near run. These findings help policymakers' policy creation. Renewable energy sources like hydro and biofuel should be prioritized above oil and coal. Carbon dioxide emissions may decrease. (Shaari et al., 2021).

Given the substantial impact of carbon dioxide (CO2) emissions on climate change and global warming, as well as the severe risks it poses to human health, the accurate forecasting of CO2 emissions is of utmost importance. The primary objective of this study is to assess the efficacy of the Inclusive Multiple Model (IMM) as an innovative methodology for forecasting carbon

dioxide (CO2) emissions within the agricultural sector of Iran. To achieve this objective, we utilised the environmental Kuznets curve (EKC) framework and collected data spanning from 2003 to 2017 for 28 provinces inside Iran. In the initial stage, distinct specifications were incorporated for the Multiple Regression (MLR), Gaussian Process Regression (GPR), and Artificial Neural Network (ANN) models. In the subsequent stage, an Integrated Model of Models (IMM) was employed to process the outputs of the most optimal specification among the Multiple Linear Regression (MLR), Gaussian Process Regression (GPR), and Artificial Neural Network (ANN) models, which were then utilised as inputs for an ANN model. The performance of the models was assessed using the Taylor diagram, as well as innovative and distinctive graphical representations. The results of the study revealed that the IMM model, with a correlation coefficient (CC) of 0.81 and a root mean square error (RMSE) of 0.69, demonstrated the most accurate estimation of CO2 emission levels. Additionally, this model exhibited the highest percentage of residuals falling within the range of -5 to 5 (37.84%), and the lowest distance from the observed data points (1.857). These enhancements provide potential avenues for future research endeavours. The use of the IMM is highly suggested for predicting CO2 emissions in order to inform the development of effective policies aimed at mitigating air pollution (Shabani et al., 2021).

This study examines China's 1990–2020 renewable energy consumption, output, export, and CO2 emissions using econometric methods. The analysis covers industry and agriculture. The study found a long-term link and causal relationship between the variables. The findings link renewable energy use, output, export, and CO2 emissions. Long-term co-integration and causality analysis show a reciprocal causal relationship between renewable energy use, output, export, and CO2 emissions. This confirms the feedback hypothesis that output and export harm the environment, while renewable energy use helps. Exports, carbon dioxide emissions, and renewable energy use are directly or indirectly linked in the short term. The growth theory is supported by this relationship. Impulse response analysis confirmed the causality test and supported the hypothesis. However, exporting industrial and agricultural goods inversely affects renewable energy use. Therefore, this relationship may hinder renewable energy sources' capacity to meet the immediate high demand for industrial and agricultural exports. Instead, a lot of fossil fuels will be used for production and exports. In order to achieve social, economic, and environmental sustainability, it is necessary to evaluate how economic growth and energy use

(both renewable and non-renewable) in important sectors affect CO2 emissions. This research also strengthens China's renewable energy sector and allows pollution control programmes to last. (Hao, 2022).

This research examines the potential correlation between the carbon dioxide (CO2) intensity of gross domestic product (GDP) and environmental degradation in Turkey, an emerging economy, from 1990 to 2018. The analysis takes into account factors such as economic growth, foreign direct investment, and renewable energy use in order to control their potential influence. This study employs various linear and nonlinear time series estimators, including the Gregory and Hansen cointegration test, bounds test, nonlinear autoregressive distributed lag (NARDL) model, fully modified ordinary least squares (FMOLS), dynamic ordinary least squares (DOLS), and canonical cointegrating regressions (CCR). The objective is to examine the potential impact of CO2 intensity of GDP, economic growth, foreign direct investment, and renewable energy consumption on environmental degradation in Turkey. The study's empirical findings indicate that the CO2 intensity of GDP plays a significant role in determining environmental deterioration in Turkey. Moreover, a decrease in the CO2 intensity of GDP is associated with a reduction in environmental degradation. Furthermore, it may be argued that economic growth plays a crucial role in determining the level of environmental sustainability in Turkey. The outcome holds significant importance for the formulation of policies and has the potential to inform and guide effective policy interventions aimed at addressing environmental challenges (Abbasi et al., 2022).

The objective of this research is to investigate the correlation between financial instability and carbon dioxide (CO2) emissions in the country of India for the period spanning from 1980 to 2020. The Autoregressive Distributed Lag (ARDL) model is employed for the purpose of assessing both long-run and short-run dynamics. Subsequently, the Vector Error Correction Model (VECM) is utilised to ascertain the causal direction. According to the results of the study, it has been determined that there is a lack of major impact of financial instability on carbon dioxide (CO2) emissions. Nevertheless, the interplay between economic development, energy consumption, and urbanisation has been found to have an adverse impact on environmental quality due to the substantial release of carbon dioxide (CO2) emissions into the surrounding ecosystem. The empirical evidence gathered in our study has substantiated the existence of an

environmental Kuznets curve. The results of the Vector Error Correction Model (VECM) indicate the presence of long-term causality in CO2 emissions, financial instability, energy use, and urbanisation. In addition, the verification of the results' validity and reliability was conducted through the utilisation of a diverse range of diagnostic tests. This study delivers original findings that contribute to the existing body of literature and may hold significant implications for policymakers in the country with regard to the financial system's role in addressing environmental issues (Qayyum et al., 2022).

The primary objective of this study is to investigate the correlation between green finance, economic growth, renewable energy consumption (specifically energy efficiency), energy import, and CO2 emissions in Vietnam. This will be accomplished through the application of multivariate time series analysis. The data were gathered between the years 1986 and 2018, coinciding with Vietnam's implementation of economic reforms known as "Doi Moi" in 1986. The study utilised the principles and techniques of cointegration, Granger causality, and error correction model (ECM) to establish the correlation between the variables under investigation. The findings of our study have provided confirmation of the presence of cointegration among the variables. The Granger causality test indicated the presence of a one-way causal relationship from renewable energy consumption to CO2 emission, as well as from green investment to CO2 emission. The findings of this study provide empirical evidence supporting the presence of cointegration among the variables. The findings of the study suggest that policies pertaining to economic development have a substantial influence on pollution levels in Vietnam. This report provides a comprehensive overview of Vietnam, focusing on its economic development, green industrial practices, environmental health, and the impact of COVID-19 on carbon dioxide emissions (Tran, 2022) .

The detrimental effects of climate change have been a focal point in numerous policy initiatives. This study examined the influence of economic development, fossil fuel usage, and population density on carbon dioxide (CO2) levels in India, Pakistan, and Bangladesh. The analysis utilised annual data spanning from 1971 to 2014. The panel Autoregressive Distributed Lags (ARDL) model was employed to estimate the long-run dynamics. Additionally, a Vector Error Correction Model (VECM) was defined to conduct a Granger causality test in order to determine the

direction of causality. The study's empirical findings have demonstrated significant connections with important policy implications, utilising a three multivariate equations model.

The initial findings from the auto-regressive distributed lags (ARDL) analysis provide support for the environmental Kuznets curve hypothesis, indicating a U-shaped link between CO2 emissions and economic development. Additionally, it can be observed that there exists a positive correlation between the consumption of fossil fuels and population density, and the subsequent release of carbon dioxide over an extended period. The empirical findings of the VECM test provide support for the existence of short-term causal relationships between economic development and CO2 emissions, population density and CO2 emissions, as well as fossil fuel usage and CO2 emissions. Additionally, it is worth noting that carbon dioxide (CO2) has been found to have adverse effects on economic development. Conversely, the long-term implications of fossil fuel usage, foreign direct investment (FDI), and total exports have been observed to have significantly favourable effects on economic development. In the near term, there is evidence to suggest that carbon dioxide emissions, fossil fuel use, and foreign direct investment (FDI) have a causal relationship with economic development, as indicated by Granger causality analysis. Finally, it should be noted that carbon dioxide (CO2) emissions have a detrimental impact on population density, but economic development has a favourable influence on population density over an extended period. Furthermore, there are short-term causal relationships between economic development and population density, as well as between CO2 emissions and population density. In the context of policy formulation, it is imperative to prioritise the use of efficient and low-carbon emission technologies (Uzair Ali et al., 2022).

The primary focus of this study endeavour has been directed on the countries of China and India. Both countries have a significant population size on a global scale, and their respective rates of economic growth have been consistently increasing annually. Nonetheless, the presence of air pollution in the form of CO2 emissions persists. Hence, the primary objective of this study is to examine the correlation between population size and gross domestic product (GDP) and carbon dioxide (CO2) emissions in China and India throughout the period from 1984 to 2014. Additionally, this research intends to offer policy suggestions pertaining to the identified issue. The data collected is analysed using the Vector Error Correction Model (VECM) in order to generate estimates. The findings of the study indicate that in China and India, both the GDP and

population have a favourable impact on CO2 emissions in both the short and long term. Additionally, the study offers policy recommendations about the readiness to pay for industry and the willingness to accept for the community. The idea of the Environmental Kuznets Curve was not validated in the context of China and India (Aminata et al., 2022).

Global economic growth depends on developing economies. Preserving environmental assets in the face of fast economic transformation remains a major obstacle for most industrialization, GDP growth, energy consumption, and urbanisation on CO2 emissions in 23 developing nations from 1995 to 2018. Our data shows that energy usage, economic growth, industrialization, and urbanisation increase CO2 emissions. A 1% increase in energy use raises CO2 emissions by 0.23%, whereas economic growth, industrialization, and urbanisation increase them by 0.17%, 0.54%, and 2.32%. Additionally, our model's short- and long-term equilibriums are adjusted annually at 0.19%. Additional robustness tests utilising DOLS and FMOLS were performed to validate the panel ARDL model's long-term results. Our analysis shows that GDP, industrial development, energy consumption, and urban expansion drive CO2 emissions in emerging nations. The panel causality study also showed a bidirectional relationship between energy consumption, GDP, urbanisation, industrialization, and CO2 emissions. This study could dramatically impact CO2 emission restrictions in selected nations. Our study can also help policymakers and stakeholders in other emerging economies make important policy decisions. Some approaches include financial incentives and infrastructural improvements that promote environmentally friendly industrialization, low-carbon technology, and sustainable urbanisation and planning. (Sikder et al., 2022) .

This study compares carbon dioxide emissions and urbanisation in the Central African Economic and Monetary Community from 1990 to 2019. Urbanisation and carbon dioxide emissions are debated in scientific literature, and current research findings are inconclusive. Carbon dioxide is the dependent variable, while GDP, urbanisation, rule of law, financial development, and government effectiveness are the independent factors. This study uses Kao and Johansen Fisher residual co-integration tests and completely modified and dynamic ordinary least squares. A statistically substantial and positive association exists between urbanisation and CO2 emissions. The causality tests show that carbon dioxide emissions affect urbanisation, GDP, and FD unit. To reduce carbon dioxide emissions, countries must improve their laws. Urbanisation laws are

needed to reduce its environmental implications, especially carbon dioxide emissions. The deployment of unregulated urbanisation practises in several nations has increased environmental impacts on people. The study suggests sustainable green urbanisation policies to promote environmental conservation through tree planting and horticulture. It's crucial to balance urban and rural development to reduce stress and congestion in state cities. Governments should encourage private sector investment in rural areas to reduce rural-urban migration. This analysis shows that CEMAC urbanisation and carbon dioxide emissions are positively correlated. The causality tests confirm the premise that carbon dioxide unidirectionally causes urbanisation, GDP, and FD. Adding value is crucial to its importance. (Ngong et al., 2022).

Globally, China, India, and the US emit the most CO2 and consume the most energy. India emits 1.80 metric tonnes of CO2 per capita, according to datacommons.org. This emission level endangers living things. This study article analyses the negative effects of CO2 emissions in India and predicts future emissions for the following decade. The projections will be based on 1980–2019 univariate time-series data. Our study used the seasonal autoregressive-integrated moving average with exogenous variables (SARIMAX), ARIMA, and Holt-Winters models. Two machine learning models—random forest and linear regression—and a deep learning-based LSTM model were also used. Multiple models help forecast data in this investigation. SARIMAX, LSTM, and Holt-Winters are the three most precise models of the six based on nine performance measures. In this study, the Long Short-Term Memory (LSTM) model predicted CO2 emissions better than others. The LSTM model had a MAPE of 3.101%, RMSE of 60.635, MedAE of 28.898, and other performance measures. Comparative research supports the same conclusion. Thus, the deep learning-based Long Short-Term Memory (LSTM) model is ideal for CO2 emission prediction. (Kumari and Singh, 2022).

The significant increase in healthcare expenses is a matter of great importance in the field of health policy on a global scale. Hence, comprehending the determinants that contribute to the rise in healthcare expenses offers policymakers empirical insights to inform their decision-making process. The objective of this study is to investigate the enduring impacts of carbon dioxide emissions, urbanisation rate, and GDP per capita on health costs. The objective of this study is to examine the impact of carbon dioxide emissions, urban population, and GDP per capita on health expenditure in a sample of 21 OECD nations from 1992 to 2018. The article

employed the Panel ARDL Approach and the Gengenbach, Urbain, and Westerlund Panel Co-integration Test. The findings of the study demonstrate the existence of a sustained correlation between health spending and many factors, including carbon dioxide emission, urban population, and GDP per capita. The impact of carbon dioxide emission (CO2), urban population, and GDP per capita on health spending is both statistically significant and positively correlated. The recent rapid economic expansion observed in OECD countries has been accompanied by a corresponding increase in environmental pollution. This has consequently led to a rise in long-term health expenditures (Kutlu and Örün, 2022).

The intent of this study is to look at how India's development affects the amount of energy and pollution each person uses. Using the Stochastic Impacts by Regression on Population, Affluence, and Technology (STIRPAT) model, this study looks at how India's energy use has changed as more people move to cities. A set of time series data from 1960 to 2015 was used for the study. Several World Bank factors were used in the study. These included population, GDP per capita, energy intensity, share of industry in GDP, share of services in GDP, total energy use, and urbanisation. These factors were used to look at the link between living in cities, having a lot of money, and using a lot of energy. There is a link between the need for energy and economic growth, especially to economic prosperity. The study also found that as cities grow, the GDP goes up, and the population grows, there will probably be a corresponding rise in carbon dioxide emissions because more energy will be used and demanded. In India, lowering the intensity of energy use is therefore a key part of improving energy security and lowering carbon dioxide pollution. The study will have big policy effects for India's energy sector as it moves towards non-conventional, clean energy sources. This is because more and more Indians are living in cities. Not many studies have looked at how population density affects the amount of energy used per person. This study also adds to the body of research on how to do these kinds of things by showing a link between how much energy each person uses, the number of people living in an area, and technology, more specifically the rates of growth in the service and manufacturing sectors. (Haldar and Sharma, 2022) .

The carbon emissions in Heilongjiang Province were measured using the IPCC calculation formula, based on the time series data from 2000 to 2019. Based on this premise, a vector autoregressive model is constructed to examine the relationship between economic growth,

industrial structure, energy intensity, energy structure, and carbon emissions. The empirical study uses the EG test and impulse response function to analyse the data. The findings indicate that a sustained equilibrium relationship exists between carbon emissions and various factors such as economic scale, industrial structure, energy intensity, and energy structure. It is observed that optimising industrial structure and energy structure can effectively mitigate the rise in carbon emissions. Conversely, the growth of economic scale and energy intensity tends to contribute to an increase in carbon emissions (Li, 2023).

The primary objective of this research endeavour is to substantiate the validity of the environmental Kuznets curve (EKC) hypothesis within the context of Indonesia, while simultaneously examining the influence of GDP per capita, income inequality, and population on carbon dioxide (CO2) emissions. The present study used a descriptive quantitative approach. The temporal data utilised in this study encompasses the time period spanning from 1990 to 2021. The employed methodology for data analysis involves the utilisation of the error correction model (ECM) to examine the impact of the independent variables on the dependent variable across both short-term and long-term periods. The findings of this study suggest that the Environmental Kuznets Curve (EKC) theory lacks empirical support in the context of Indonesia, both in the short-term and long-term. The relationship between GDP per capita and CO2 emissions exhibits a notable beneficial impact, which is observed in both the immediate and extended periods. In the short run, income disparity does not exhibit a statistically significant positive impact on CO2 emissions, while in the long term, it does not have a statistically significant negative impact. In the near run, the population does not exhibit a major adverse impact on CO2 emissions, while in the long term, it demonstrates a noteworthy positive influence (Yunita et al., 2023).

Developing nations aspire to achieve industrialization and sustainable growth, frequently disregarding the potential environmental repercussions. Nevertheless, there is a scarcity of empirical research that has examined the impact of economic transition led by industrialization on carbon footprint in developing countries, particularly using a non-parametric methodology. This study examines the influence of industrial value-added (IGVA), population, energy intensity, and carbon intensity on CO2 emissions in India, utilising Kaya's identity and the novel multivariate quantile-on-quantile regression (QQR) model, which extends Sim and Zhou's (2015)

bivariate QQR model. The research is conducted within a specific context. This study is among the initial contributions in the existing body of literature that examines the relationship between industrialization and carbonization within the framework of Kaya's identification for the Indian economy. It employs a novel multivariate Quantile Quantile Regression (QQR) approach, thereby making a valuable methodological and empirical contribution to the existing literature. The results of the multivariate quantile regression technique indicate that achieving economic and environmental development necessitates the implementation of persistent long-term plans. The empirical data indicate that there is no evidence to support the claim that India's carbon emissions have increased as a result of its industrialization. This suggests that the Industrial Gross Value Added (IGVA) is negatively and significantly associated with CO2 emissions. In some quantiles, there is a positive correlation between population size and CO2 emissions. In contrast, the carbonization process within the Indian economy is subject to asymmetrical influences from factors such as GDP per capita, energy intensity, and carbon intensity. The findings of the quantile Granger causality analysis provided additional support for the results presented earlier. The present analysis additionally provides policy recommendations aimed at promoting environmentally sustainable economic growth and attaining India's sustainable development goals (SDGs) (Das et al., 2023).

2.4 Stationary Test

A stationary series is characterised by statistical features, such as mean, variance, covariance, and standard deviation, that remain constant across time. In other words, these statistical properties do not exhibit any temporal variation or dependence. In essence, the concept of stationarity in Time Series analysis refers to a series that lacks both Trend and Seasonal components. Statistical models are more effective and exact in predicting stationary series.

2.4.1 Augmented Dickey - Fuller Test

Statistical tests are predicated upon robust assumptions regarding the characteristics of the data. The sole purpose of their utilisation is to provide information regarding the extent to which a null

hypothesis can be either rejected or not rejected. In order for a given problem to possess meaningful interpretation, the result must be appropriately analysed. Nevertheless, these tests offer a rapid means of verifying and validating whether the time series exhibits stationarity or non-stationarity. The Augmented Dickey-Fuller test is classified as a unit root test, which falls under the category of statistical tests. In the field of probability theory and statistics, the concept of a unit root pertains to certain stochastic processes, such as random walks, and can give rise to challenges when performing statistical inference within the context of time series models. In more precise words, the concept of a unit root refers to a non-stationary time series that may or may not have a trend component. The ADF test is performed on the following assumptions:

- The null hypothesis (H0) posits that the series under consideration is non-stationary, meaning it possesses a unit root.
- The alternate hypothesis (HA) posits that the series under consideration is stationary, meaning that it does not possess a unit root.
- If the null hypothesis is not rejected, this test may offer evidence indicating that the series is non-stationary.

There are several conditions that may lead to the rejection of the null hypothesis (H0). If the test statistic is less than the critical value and the p-value is less than 0.05, the null hypothesis (H0) is rejected. This implies that the time series does not exhibit a unit root and can be considered stationary. The structure of the phenomenon under consideration is not contingent upon time.

2.5 Causality Test

2.5.1 Granger Causality Test

The proposed methodology involves employing an econometric hypothetical test to assess the extent to which one variable can be utilised to predict another variable in the context of multivariate time series data while considering a specific time lag. The idea of Granger causality is a statistical tool employed to assess the predictive relationship between two time series, specifically whether one time series may be used to forecast future values of another time series. The measure quantifies the degree to which historical values of one variable offer meaningful

insights into predicting the future behaviour of another variable. The application of Granger causality facilitates the examination of probable associations among variables, hence assisting in the prediction of trends. Nevertheless, it is crucial to acknowledge that Granger causality does not establish direct causation and should be approached with caution when assessing causal connections between variables. In order to do the Granger Causality test, it is necessary for the data to exhibit stationarity, which implies a consistent mean, constant variance, and the absence of any seasonal patterns. The non-stationary data can be converted into stationary data by applying differencing techniques, such as first-order or second-order differencing. It is advisable to abstain from conducting the Granger causality test if the data does not exhibit stationarity even after undergoing second-order differencing.

2.6 Introduction to Time Series Forecasting

The discipline of data analysis encompasses a notable area of research known as time series forecasting, which involves the prediction of future values by analysing patterns in historical data. From the forecasting of meteorological conditions to the analysis of financial markets and the prediction of sales trends, this methodology offers important knowledge for the purpose of making informed decisions. Through the analysis and modelling of temporal data, time series forecasting provides us with the capacity to reveal inherent trends, seasonality, and patterns, so enabling us to provide well-informed forecasts regarding future outcomes.

Time series data is comprised of observations that are collected in a consecutive manner over a period of time. The collection of these data points is commonly conducted at regular intervals, including daily, weekly, monthly, or yearly. The main aim of time series forecasting is to make predictions about future values or events by analysing patterns found in past data. Through the application of statistical tools and mathematical models, analysts are able to extract significant information, identify interrelationships, and forecast prospective patterns.

2.6.1 Importance of Time Series Forecasting

Time series forecasting is of significant importance in diverse areas, as it has a profound impact on decision-making processes and facilitates the improvement of operational efficiency.

There are other noteworthy applications that might be mentioned.

- Economics and finance disciplines extensively utilize time series forecasting techniques to make predictions in financial markets. These forecasts are crucial for several purposes such as forecasting stock market movements, predicting asset prices, modelling exchange rates, and analysing economic trends. These prognostications aid investors, governments, and financial institutions in formulating well-informed judgements and mitigating risks.
- Sales and demand forecasting is a crucial aspect of business operations, as it allows organisations to predict future demand patterns and effectively manage their inventory. Time series forecasting is a commonly employed method in this regard. By utilising this approach, businesses can make informed decisions regarding inventory levels and ensure optimal management of their resources. Through the comprehension of customer behaviour and analysis of previous sales data, organisations have the power to adapt their production, marketing, and supply chain strategies, resulting in enhanced profitability.
- Weather and climate prediction involves the utilisation of time series forecasting techniques by meteorologists to provide predictions about future weather conditions and long-term climate trends. These projections play a crucial role in enhancing disaster preparedness, facilitating farm planning, optimising energy management, and facilitating efficient resource allocation.
- Healthcare and disease outbreaks: Time series analysis facilitates the anticipation of disease outbreaks, enabling public health authorities to efficiently allocate resources, conduct preventive measures, and swiftly respond to emergencies. Furthermore, healthcare practitioners utilise forecasting techniques to predict patient admissions, strategize staff schedules, and optimise the allocation of resources.

2.6.2 Different types of Time Series Techniques

Time series forecasting employs several methodologies and models, which are selected based on the specific properties of the data and the desired level of accuracy. Several often-utilised strategies include:

- Moving averages are a statistical tool used to reduce irregularities and emphasise underlying patterns by computing the mean of a predetermined number of previous observations.
- The utilisation of Simple Moving Average (SMA) and Exponential Moving Average (EMA) is prevalent within this particular domain.
- The Autoregressive Integrated Moving Average (ARIMA) models are capable of capturing both autoregressive and moving average components inside a given time series. By employing the technique of differencing to achieve stationarity in the data, it becomes possible to utilise ARIMA models for the purpose of forecasting future values.
- The Seasonal Decomposition of Time Series (STL) method is employed to partition a time series into distinct components, namely trend, seasonality, and residual. This enables analysts to examine and predict each component individually. This method is particularly advantageous when applied to data that exhibits pronounced seasonal patterns.
- Machine learning methodologies, such as neural networks, support vector regression (SVR), and random forests, have become increasingly prominent in the field of time series forecasting. These models have the ability to capture intricate connections and non-linear patterns, frequently exhibiting superior performance compared to conventional statistical methods.
- A multivariate time series is characterised by the presence of many time series variables. Each variable exhibits a dependence not just on its own previous values but also on the values of other variables. The aforementioned dependency is utilised for the purpose of predicting forthcoming values. Vector Auto Regression (VAR) is a widely employed technique for forecasting multivariate time series. Other deep learning methods include utilising a combination of several historical time series alongside Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM) networks, to generate forecasts pertaining to future events.

2.6.3 Different Types of Challenges and Considerations

Time series forecasting is a field with significant potential, but it also entails various obstacles that necessitate careful consideration.

- Seasonality and trend variations pose a significant challenge in effectively modelling time series data. In order to prevent inaccurate predictions, analysts are required to recognize and include these patterns into their analysis.
- The accuracy of time series projections can be severely affected by incomplete or erroneous data, specifically in terms of data quality and missing values. Ensuring trustworthy results necessitates the implementation of appropriate data cleaning, imputation techniques, and outlier detection approaches.
- The accuracy of time series forecasts diminishes as the forecast horizon stretches further into the future. Analysts are required to achieve a harmonious equilibrium between short-term and long-term prognostications, taking into account the accessible data and the desired degree of accuracy.
- Volatility and uncertainty pose challenges for time series forecasting when unforeseen occurrences or external influences disrupt the distribution of data. The issues can be mitigated by adapting models to accommodate volatility and adding external inputs.

2.7 Use cases of Time Series Forecasting in prediction of CO2 emissions

Time series forecasting plays a crucial role in predicting CO2 emissions, as it helps governments, organizations, and individuals make informed decisions to mitigate climate change and promote sustainability. Here are some important use cases of time series forecasting in predicting CO2 emissions:

1. Policy Planning and Compliance Monitoring:
    - Governments and regulatory bodies can use time series forecasting to estimate future CO2 emissions and set emission reduction targets. These forecasts help in designing and implementing policies to curb emissions effectively.

- Monitoring compliance with emission reduction agreements, such as the Paris Agreement, by comparing actual emissions with forecasted emissions.

2. Energy Production and Consumption Optimization:
    - Utilities and energy companies can use time series forecasting to predict energy demand and optimize power generation, thus reducing CO2 emissions associated with excessive energy production.
    - Consumers can make informed decisions about when to use energy-intensive appliances or shift their energy consumption to times when cleaner energy sources are more abundant.
3. Transportation and Traffic Management:
    - Forecasting CO2 emissions from transportation can aid in optimizing traffic management systems, promoting the use of public transportation, and encouraging the adoption of electric vehicles.
    - Predicting peak traffic times and congestion can help reduce idling and fuel consumption, leading to lower emissions.
4. Supply Chain and Inventory Management:
    - Businesses can use CO2 emission forecasts to optimize their supply chains and inventory management. Reducing unnecessary transportation and storage helps lower emissions.
    - Companies can choose suppliers and transportation methods that have lower carbon footprints based on predictive emission data.
5. Renewable Energy Integration:
    - Forecasting CO2 emissions enables better integration of renewable energy sources like wind and solar power into the grid. Accurate predictions help grid operators balance supply and demand efficiently.
    - Predicting periods of high carbon intensity (e.g., during fossil fuel-heavy energy generation) allows consumers to adjust their energy usage.
6. Emission Reduction Initiatives:
    - NGOs and environmental organizations can use forecasting to assess the effectiveness of emission reduction initiatives over time. They can adjust their strategies based on whether emissions are trending up or down.

- Funders and investors in sustainable projects can make data-driven decisions on which initiatives to support.

7. Building and Urban Planning:
    - Urban planners can use CO2 emission forecasts to design more sustainable cities. This includes optimizing public transportation routes, encouraging green building practices, and reducing the carbon footprint of urban areas.
    - Predicting the impact of new developments on emissions helps authorities make informed zoning and building code decisions.
8. Carbon Pricing and Trading:
    - Predicting future CO2 emissions assists in setting appropriate carbon pricing and trading mechanisms. It helps companies make decisions about whether to buy carbon credits or invest in emission reduction technologies.
9. Environmental Impact Assessment:
    - Time series forecasting aids in assessing the environmental impact of various activities, such as land use changes, deforestation, and industrial expansion. It allows for better planning to mitigate potential emissions.
10. Public Awareness and Education:
    - Forecasted CO2 emissions can be used in educational campaigns to raise public awareness about the consequences of different activities on the environment. This can lead to more environmentally responsible behaviour.

Time series forecasting in CO2 emissions prediction is an essential tool for addressing climate change and making data-driven decisions to reduce greenhouse gas emissions across various sectors of society.

2.8 Summary

This chapter introduces greenhouse gases, focusing on CO2 emissions, primarily from burning fossil fuels in residential and commercial properties. It examines the relationship between GDP, population, and CO2 emissions using various modeling methods and identifies research gaps. The sub-chapter discusses the importance of Time Series Forecasting, various methods, challenges, and use cases in predicting CO2 emissions, helping governments and organizations make policy changes and mitigate climate change.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

The study necessitates a dataset which encompasses historical data related to mainly CO2 and Greenhouse gas emissions of India. The first step in this process entails obtaining the suitable dataset and doing a comprehensive analysis to construct a resilient model that produces the intended results. The initial subsection of the chapter delineates the research approach, furnishing a comprehensive summary of the overarching procedures and strategy. Furthermore, this paper examines the procedure of dataset selection, encompassing relevant information pertaining to the selected dataset. The chapter outlines the essential pre-processing and cleaning strategies that should be done to enhance the outcomes, taking into account the dataset. Following this, the chapter explores the selection of algorithms for time series prediction and provides an overview of the corresponding procedures involved in this process.

Moreover, the chapter provides a detailed explanation of the architectural design of the models that will be employed. Finally, this chapter introduces performance evaluation metrics, which are essential for evaluating the precision and efficacy of the models.

3.2 Research Approach

After doing extensive literature review of various approaches and methodologies by many researchers, the research for this study is mentioned in this chapter. Identification of the suitable dataset required for the study is one of the preliminary steps of this study. The pre-processing techniques, missing values imputation using mean, median have been explored and their effect on the model has been examined.

The approach involves steps such as Data Collection, Data pre-processing, Exploratory Data analysis. Further, the study explores Vector Auto regression Time Series analysis and Multiple Linear regression modeling methods by training, testing and evaluation of both the methods.

Figure 3 shows the framework with the workflow of both models. In the next section, we will provide you with an overview of the dataset.

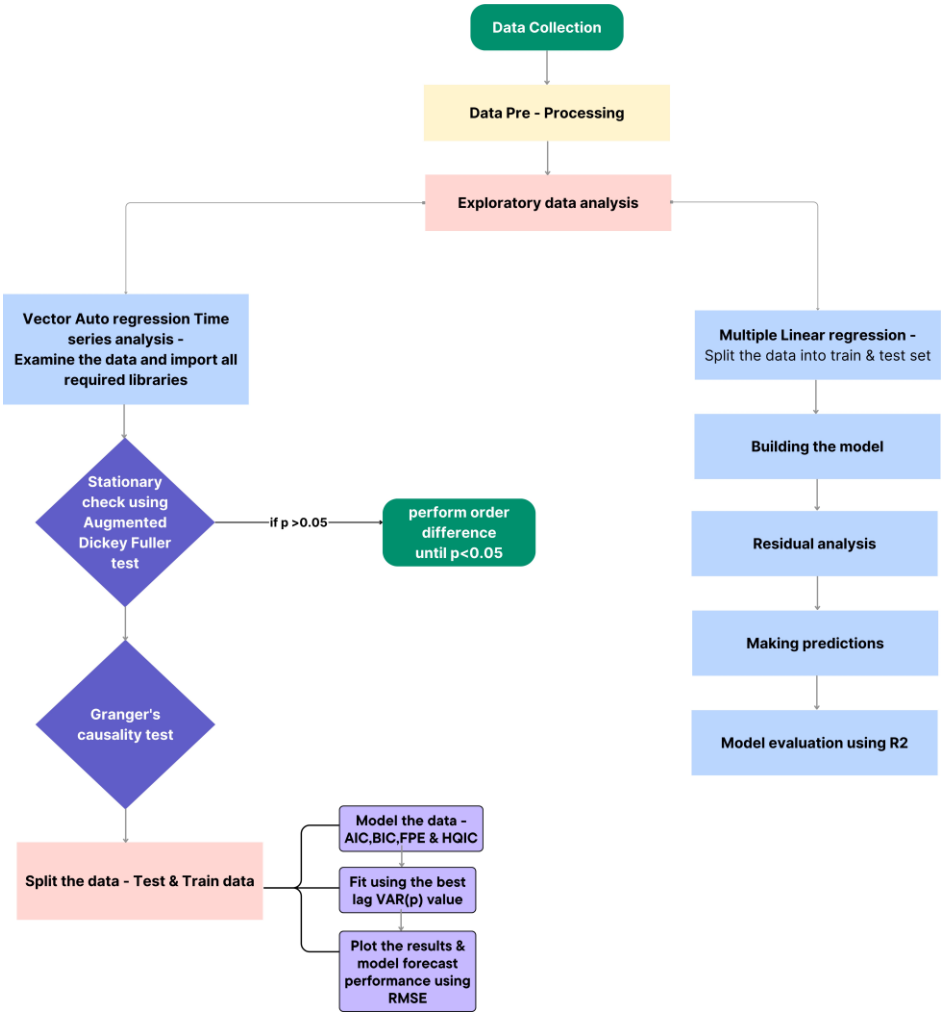


Figure. 3 VAR & Multiple linear regression Framework illustrating workflow

3.3 Dataset Description

The dataset, available on the Our World in Data platform, contains 1,705,669 observations with 79 attributes. Among these attributes, three are relevant to the topic: Country, GDP, population, and co2_including_luc. The data covers historical information from 1850 to 2021 for these attributes. Specifically, we are focusing on data related to India for conducting a time series analysis. However, it is important to note that the GDP column has 19% missing data, and the co2_including_luc column has 11% missing data. These missing values can be suitably imputed to facilitate further processing and analysis of the data.

year	country	population	gdp	co2_including_luc
1850	India	235732435.0	2.230000e+11	531.207
1851	India	236722812.0	2.240000e+11	531.207
1852	India	237693556.0	1.258971e+12	531.207
1853	India	238644470.0	1.258971e+12	531.207
1854	India	239599135.0	1.258971e+12	531.207

Figure. 3.1 Head of Data pertaining to India – gdp, population and co2_including_luc

The dataset can be broadly classified into five sections:

- **CO2 emissions**: What is the annual quantity of carbon dioxide emissions produced by a given country? What is the mean level of emissions per individual? What is the cumulative amount of emissions it has released over a given period? What is the comparative analysis of emissions after accounting for trade adjustments?
- **Coal, oil, gas, cement**: The quantification of carbon dioxide emissions originating from several sources, namely coal, oil, gas, flaring, and cement manufacture, is of significant interest.
- **Other greenhouse gases**: What is the aggregate quantity of greenhouse gases emitted by each country? What are the quantities of methane and nitrous oxide emissions?
- **Emissions by sector**: In terms of emissions, it is necessary to inquire about the industries that provide the most significant contributions. To what extent does the transportation sector contribute in comparison to the electrical sector? Additionally, what is the magnitude of emissions from the agriculture and land use sectors?
- **Carbon and energy efficiency**: The topic of inquiry pertains to the quantification of energy consumption relative to gross domestic product (GDP) in the context of carbon and energy efficiency. What is the quantity of carbon emissions produced per unit of energy?

3.4 Pre-Processing and Exploratory Data Analysis

The data will be processed, and the following methods will be employed accordingly.

- **Drop columns**: Drop the unwanted columns from the dataset excluding country, year, population, gdp, co2_including_luc pertaining to India.
- **Handling Missing Values**: We can see the presence of missing values in the data. The 19% of GDP missing values and 11% of co2_including_luc missing values will be imputed with mean median or mode values respectively. The changed dataset will be used for further analysis.

year	country	population	gdp	co2_including_luc
1850	India	235732435.0	2.230000e+11	NaN
1851	India	236722812.0	2.240000e+11	NaN
1852	India	237693556.0	NaN	NaN
1853	India	238644470.0	NaN	NaN
1854	India	239599135.0	NaN	NaN

Figure. 3.2 Missing values in the dataset

- **Outlier Analysis**: In addition to doing an assessment of missing values within the dataset, it is imperative to also examine the presence of outliers. This is crucial as outliers have the potential to substantially impact the statistical characteristics of a time series, hence resulting in erroneous conclusions and subpar forecasting accuracy. Outliers may arise due to errors in data collection or may serve as indicators of variability within the dataset. The methodology employed in this study will involve the utilisation of box plot analysis. The objective of outlier identification is to identify patterns within data that deviate from anticipated behaviour. The significance of these entities in the realm of data stems from their ability to convert into practical information across a diverse range of applications (Singh and Upadhyaya, 2012). After handling the missing values and outlier plot analysis, the correlation between the three variables gdp, population and co2_including_luc is plotted using the heat map plot.

3.5 Multivariate Time Series Analysis

The subsequent techniques outline the implementation of multivariate time series analysis using the Python programming language.

- **Vector Autoregression (VAR)**: The Vector Autoregression (VAR) model is a statistical technique used to analyse the dynamic relationship between multiple time series variables. The Vector Autoregression (VAR) approach employs an Autoregressive (AR) model. The concept being discussed pertains to the extension of autoregression (AR) to encompass several parallel time series.
- **Vector Autoregression Moving-Average (VARMA)**: The Vector Autoregression Moving-Average (VARMA) model is a statistical method used in time series analysis. The concept being discussed pertains to the extension of ARMA models to encompass several parallel time series, specifically multivariate time series.
- **Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX)**: The Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX) model is a statistical method used in econometrics to analyse time series data. The Vector Autoregression Moving-Average with Exogenous Regressors (VARMAX) model is a variant of the VARMA model that incorporates the incorporation of exogenous variables into the modelling process. The method being referred to is a multivariate extension of the ARMAX technique.
- **Holt Winter's Exponential Smoothing (HWES)**: Holt Winter's Exponential Smoothing (HWES) is a forecasting method that utilises exponential smoothing techniques. The Holt Winter's Exponential Smoothing (HWES) method is a mathematical model that incorporates exponential weighting to predict future values based on past observations. It considers both trends and seasonality patterns in the data.

The Vector Autoregression model will be employed for analysing, training, testing and evaluation of the data.

3.6 Vector Auto Regression (VAR)

The general form of a vector autoregression (VAR) equation is:

$$y\_t = \beta 0 + \beta 1\, y\_\{t-1\} + \beta 2\, y\_\{t-2\} + \ldots + \beta p\, y\_\{t-p\} + \gamma 1\, x\_\{t-1\} + \gamma 2\, x\_\{t-2\} + \ldots + \gamma q\, x\_\{t-q\} + \epsilon t$$

where:

- yt is a vector of endogenous variables at time t.
- $\beta 0$ is the constant term.
- βi are the coefficients for the lagged values of yt.
- γj are the coefficients for the lagged values of xt, which are exogenous variables.
- ϵt is the error term.

The VAR model can describe the relationship between many quantities over time. The number of endogenous variable lags depends on the VAR model order, p. VAR models with higher orders are more complex and require more data to estimate.

The VAR model has multiple applications, such as variable forecasting and causal connection analysis, identifying system shocks and Understanding system dynamics. The VAR model excels at multivariate time series analysis. It can help you grasp complex interactions between variables and predict future values.

In general, VAR modeling involves the following steps:

- Analyse the Time Series Characteristics.
- A test to establish causation within a time series.
- Assess the stationarity of the data.
- If necessary, convert the series into a stationary form.
- In order to determine the optimal order (p), it is necessary to do an analysis.
- Create training and test datasets in order to facilitate the process of training and evaluating machine learning models.
- Conduct training on the model.
- Reverse the transformation, if applicable.
- The model's performance will be assessed by employing the test set.
- Predicting the future

3.6.1 Visualize and Analyse Time Series Data

The three variables – gdp, population and co2_including_luc is visualized by using the Line chart. The visualization is done to observe the trend and seasonality of the data.

3.6.2 Check for Stationarity

Prior to using the VAR model, it is imperative to ascertain the stationarity of all variables within the dataset. The concept of stationarity pertains to a statistical characteristic in which a series demonstrates a consistent mean and variance throughout its duration. The Augmented Dickey-Fuller (ADF) test is a frequently used approach for evaluating stationarity.

3.6.2.1 Unit Root Test

The presence of a unit root in a time series is a defining feature that renders it non-stationary. The ADF test is classified as a unit root test. The presence of a unit root in a time series is shown when the value of alpha equals 1 in the equation provided.

$$\boldsymbol{Yt} = \propto \boldsymbol{Yt} - \mathbf{1} + \boldsymbol{\beta Xe} + \in$$

where Yt is value of the time series at time 't' and Xe is an exogenous variable.

The existence of a unit root implies that the time series exhibits non-stationarity. The Dickey-Fuller test is a statistical test used to determine the presence of a unit root in a time series. It specifically examines the null hypothesis that the coefficient α in the model equation is equal to 1. The symbol α (alpha) represents the coefficient of the initial lag on the variable Y. The Dickey Fuller equation is represented below.

Null Hypothesis (Ho): α (alpha) =1

$$\boldsymbol{y_t = c + \beta t + \alpha y_{t-1} + \Phi \Delta Y_{t-1} + e_t}$$

where, Y(t-1) = lag 1 af time series and ø(delta) Y(t-1) is first difference of time series at time(t-1).

The null hypothesis of this test is fundamentally identical to that of the unit root test. The coefficient of Y(t-1) is 1, indicating the existence of a unit root. If the series is not rejected, it is considered to be non-stationary. The Augmented Dickey Fuller Test is derived from the equation mentioned above and is widely recognised as a prevalent method for doing Unit Root tests.

The ADF test can be considered as an augmented version of the Dicker Fuller test. The Augmented Dickey-Fuller (ADF) test extends the Dickey-Fuller test equation by incorporating a higher order of autoregressive process within the model. The Augmented Dickey Fuller equation is represented as below.

$$y_t = c + \beta t + \alpha y_{t-1} + \Phi \Delta Y_{t-1} + \Phi \Delta Y_{t-2} .. + \phi_p \Delta Y_{t-p} + e_t$$

In this experiment, the null hypothesis posits that the time series exhibits non-stationarity. When the p-value obtained from the Augmented Dickey-Fuller (ADF) test is below the significance level of 0.05, the null hypothesis is rejected. This implies that the time series variables under consideration are deemed to be stationary. To attain stationarity, it is necessary to iteratively use the differencing operation until the data frame exhibits stationarity.

3.6.3 Granger's Causality Test

This is essentially an economic hypothetical test that aims to validate the inclusion of a specific variable in the forecasting of another variable inside a multivariate time series dataset, while considering a specific lag.

The concept of Granger causality is a statistical tool employed to ascertain the predictive relationship between two time series variables. The measure quantifies the degree to which historical values of a given variable offer meaningful insights into predicting the future behaviour of another variable. The concept of Granger causality is employed to examine potential associations between variables, hence facilitating the prediction of trends. Nevertheless, it is crucial to acknowledge that Granger causality does not establish direct causation and should be approached with caution when assessing causal connections between variables. In order to do the Granger Causality test, it is necessary for the data to exhibit stationarity, which implies a consistent mean, constant variance, and the absence of any seasonal patterns.

Granger's causality test determines variable relationships before modelling. If the p-value of the statistical test is found to be less than 0.05, it can be concluded that there exists a statistically significant relationship between the variables under investigation. Conversely, if the p-value exceeds 0.05, it can be inferred that there is no statistically significant association between the variables. Granger's causality test allows for the identification of both unidirectional and bidirectional causality between the variables.

3.7 Split & Model the Data

Split the dataset into train data (95%) and test data (5%) to model it. Input into the VAR module for modelling purposes. After this, the next step is to select the best-fit lag value. To do this we need to compare different AIC (Akaike information criterion), BIC (Bayesian information criterion), FPE (forecast prediction error) & HQIC (Hannan Quinn Information Criterion). The above parameters will help in the selection of the best-fit lag value.

- **AIC (Akaike information criterion)**: The Akaike information criterion (AIC) is a metric used to assess the quality of a model by considering both its goodness of fit to the observed data and its level of complexity. The Akaike Information Criterion (AIC) offers a method for evaluating and selecting among many models, aiming to identify the model that exhibits the optimal fit to the data while maintaining the lowest level of complexity.

- **BIC (Bayesian information criterion)**: The Bayesian information criterion (BIC) is a statistical measure used in model selection. This statistical metric is employed for the purpose of selecting a model from a collection of potential models. Similar to the Akaike information criterion (AIC), the Bayesian information criterion (BIC) offers a balance between the quality of fit and the intricacy of the model. Nevertheless, the Bayesian Information Criterion (BIC) imposes a more stringent penalty on the quantity of parameters compared to the Akaike Information Criterion (AIC), hence aiding in the mitigation of overfitting.

- **FPE (forecast prediction error)**: Forecast error, sometimes referred to as prediction error or forecasting error, is a metric that evaluates the precision of a forecast or prediction in relation to the observed event. The calculation of forecast error involves determining the discrepancy between the anticipated value and the factual value within a specified time frame or observation. The forecast error may have a positive or negative number, contingent upon whether the forecast surpasses or falls short of the actual value. A positive prediction error signifies an underestimation in the forecast, whereas a negative forecast error implies an overestimation in the forecast.

- **HQIC (Hannan Quinn Information Criterion)**: The Hannan-Quinn information criterion (HQC) is a widely employed metric for evaluating the adequacy of a statistical model's fit. It is commonly utilised as a means of selecting the most appropriate model from a limited number of options. The method in question does not rely on the log-likelihood function (LLF), but rather has a connection to Akaike's information criterion.
  Like the Akaike Information Criterion (AIC), the Hannan-Quinn Criterion (HQC) incorporates a penalty factor for the number of parameters in the model. However, the penalty term in HQC is greater than that in AIC.

3.8 Fit the data using the best lag value

In order to determine the appropriate order for the VAR model, a method of iterative fitting is employed. This involves fitting VAR models of increasing orders and afterwards selecting the order that yields a model with the lowest Akaike Information Criterion (AIC). The minimum values in combination with AIC, BIC, FPE, and HQIC are given the '*' sign. Select the minimum lag value VAR(p) & fit the model accordingly. Co-efficient, Std. error, t-stat and model probabilities can be seen at every lag till the best-fit lag value.

3.9 Forecast, predict & plot the results

The generated projections are limited to the scale of the training data employed by the algorithm. In order to restore the data to its original scale, it is necessary to reverse the differencing process the same number of times as the original input data was differenced. In this particular instance, the occurrence is twice. The forecasts are rescaled to their original magnitude. Subsequently, the

forecasts are graphically shown in relation to the actual values obtained from the test dataset.

3.10 Evaluation

The predicted versus original values of the variables for test data and compare the same to check the accuracy of the forecasted data and actual values metric Root Mean Square error (RMSE), can be used.

3.10.1 Root Mean Square Error

The error term derived by mean square error (MSE) is not in the same dimension as the target variable 'y' due to its squared nature. To address this, the root mean square error (RMSE) metric is employed, which involves taking the square root of the MSE value.

$$RMSE = \sqrt{[(1/n) * \Sigma (y_i - \hat{y}_i)^2]}$$

Where:

RMSE: Root Mean Squared Error

n: The number of data points (samples)

y_i: The actual (observed) value of the i-th data point

$\hat{y}_i$: The predicted value of the i-th data point

3.11 Multiple linear regression

Multiple linear regression is a statistical method used to analyze the correlation between a single dependent variable and multiple independent variables (also known as explanatory variables). The primary aim of multiple regression is to identify a linear equation that can most accurately predict the value of the dependent variable Y based on various values of the independent variables in X.

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + ... + \beta nxn + \varepsilon$$ (Linear Regression: Introduction to Multiple Regression - codingstreets, 2023)

In this equation:

- y stands for the dependent variable (the variable we want to predict).
- β0 represents the intercept (the value of y when all independent variables are zero).
- β1, β2, β3, ..., βn are the coefficients (indicating how much y changes for a one-unit change in each corresponding independent variable x1, x2, x3, ..., xn).
- x1, x2, x3, ..., xn refer to the independent variables.
- ε denotes the error term, which captures the difference between the actual and predicted values of y.

3.12 Data Preparation

3.12.1 Handling Categorical Variables

When dealing with several variables, it is possible that certain category variables may prove to be valuable for the model. Therefore, it is imperative to effectively manage these variables in order to obtain a robust model. One potential approach for addressing this issue involves the utilisation of dummy variables. The fundamental concept underlying the creation of dummy variables is the generation of additional columns to represent categorical variables with 'n' levels. Each of these new columns signifies the presence or absence of a certain level through the use of binary values, namely zero or one. It is important to note that ‘n-1’ dummy variables are created to avoid multicollinearity. A dummy variable is created for ‘country’ variable.

3.13 Spilt the data into the train & test set

The initial fundamental step in regression is to carry out a train-test (70-30) split. To accomplish this, we employ the sci-kit-learn model. Proper feature scaling plays a crucial role when dealing with numerous independent variables in a model. Since many of these variables may be on different scales, it can lead to obscure coefficients that are difficult to interpret.

3.13.1 Feature Scaling

Scaling the features using the MinMax scaler is essential for two reasons: it enhances the ease of

interpretation and facilitates faster convergence of gradient descent methods. With the MinMax scaler, the variables are transformed in a way that all values fall between zero and one, using the maximum and minimum values in the data.

3.13.1.1 Min Max Scaling

The variables are normalized to a range of zero to one by utilising the minimum and maximum values present in the dataset.

$$\textit{Scaled Value } (X\_scaled) = (X - X\_min) / (X\_max - X\_min)$$

Where:

X: The original value of the data point for a particular feature.

X_min: The minimum value of the feature in the dataset.

X_max: The maximum value of the feature in the dataset.

3.14 Building the Model

The initial crucial step in constructing the model involves dividing the training data into separate sets, X, and y. Subsequently, the Recursive Feature Elimination (RFE) and linear regression modules from sci-kit-learn are imported. RFE is executed with the desired number of output variables. The process then entails creating the X_test data frame with the RFE-selected variables. Afterwards, a constant variable is added, and the linear model is run using the Ordinary Least Squares method. Features that have a high p-value, indicating that they are not significantly impactful in predicting the dependent variable, are dropped. In regression model building, the null hypothesis corresponding to each p-value posits that the associated independent variable does not influence the dependent variable, while the alternate hypothesis states that it does impact the response. A low p-value (less than 0.05) suggests that the null hypothesis can be rejected.

3.14.1 Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) is a statistical measure that addresses the limitation of relying just on correlations. It recognizes that a single variable may not fully account for the variation in another variable, but a combination of variables may collectively provide a more comprehensive explanation. In order to assess the relationships between variables, the Variance

Inflation Factor (VIF) is employed. The Variance Inflation Factor (VIF) is a statistical measure that aids in elucidating the link between one independent variable and all other independent variables. The equation for the Variance Inflation Factor (VIF) is presented as follows:

$$VIF = 1 / (1 - R^2)$$

Where:

- VIF: Variance Inflation Factor for a specific independent variable.
- R^2: The coefficient of determination of a regression model where the independent variable in question is regressed against all the other independent variables in the model.

Additionally, redundant features are identified using correlations and Variance Inflation Factor (VIF) >5 in conjunction with p value are subsequently dropped. The model is then rebuilt, and the process is repeated iteratively.

3.15 Residual analysis

To perform a Residual Analysis on the train data, we aim to verify whether the error terms exhibit a normal distribution, which is a fundamental assumption in linear regression. To achieve this, we plot a histogram of the error terms and visually examine their distribution. This step is essential in validating the key assumptions of linear regression and ensuring the reliability of the model.

3.16 Making Predictions

After applying scaling on the test sets, we divide them into X_test and y_test. We then utilize our model to make predictions. To do this, we create a new data frame called X_test_new by dropping certain variables from X_test. After adding a constant variable, we proceed to make predictions using the model.

3.17 Model evaluation

The model evaluation involves two steps. First, we visually assess the performance by plotting the actual y_test values against the predicted y_predicted values. This comparison allows us to

observe how well the model's predictions align with the actual data. Second, we utilize the R^2 value to evaluate the model's performance on the test data. The R^2 value provides a measure of how well the model fits the test data and indicates the proportion of the variance in the dependent variable that can be explained by the independent variables.

3.18 Summary

This chapter provides an overview of the proposed research approach. The study begins with a concise introduction that offers an overview of the research and presents the dataset that will be employed in this investigation. The methodology is subsequently explicated using a flowchart, which visually demonstrates the sequential approach. Furthermore, the chapter also explores preprocessing strategies that are designed to enhance the performance of the model. The subject matter also encompasses a range of components related to time series data, such as examinations of segments, stationarity, and autocorrelation. Different methodologies for time series forecasting are also investigated. Additionally, the chapter discusses the proposed model and its potential benefits for this thesis. We also engaged in a discussion regarding the metrics of assessment that will be utilised to evaluate the success of the model. The metrics will have a significant impact on assessing the efficacy and precision of the planned research.

CHAPTER 4: IMPLEMENTATION AND ANALYSIS

4.1 Introduction

This section elucidates the practical implementation of the research approach outlined in this study with respect to the chosen dataset. This would involve providing a detailed explanation of the data preparation process, including the treatment of missing values, the handling of outliers, and the management of highly correlated features. The initial step involves the examination and exploration of the data through various visual representations, such as data distribution analysis. Next, we proceed to employ pre-processing techniques to effectively cleanse the data, hence enhancing the accuracy of predictions generated by the model. Next, we proceed to do exploratory data analysis through the creation of several graphs. These visualizations aid us in gaining a deeper knowledge of the inherent characteristics of the dataset. Finally, we conclude this section by providing an interpretation and evaluation of the model.

4.2 Dataset Description

The dataset, available on the Our World in Data platform, contains 1,705,669 observations with 79 attributes. Among these attributes, three are relevant to the topic: gdp, population, and co2_including_luc. The data covers historical information from 1850 to 2021 for these attributes. Specifically, we are focusing on data related to India for conducting a time series analysis.

4.3. Dataset Preparation and Exploration

We will be performing the following steps:

- **Reading and understanding the data**: Import the required libraries and load the downloaded dataset into python notebook.
- **Initial Inspection**: Check the structure of the dataset – information, size, and shape.
- **Data Description**: Understand the meaning and context of each column in the dataset by referring to the data documentation of dataset.

- **Data Summary**: Generate summary statistics to get an overview of the data, including mean, median, standard deviation, minimum, maximum, etc.
- **Data quality and Types**: Check the data types of each column to ensure they are appropriate for analysis. Convert columns to the correct data types if necessary. Check the data quality and missing values.

Read the dataset with index column as ‘year’. Check the information, shape, and size of the dataset. The dataset has 50598 rows and 78 columns. Specifically, we are focusing on data related to India for conducting a time series analysis with only three variables – ‘gdp’, ‘population’ and ‘co2_including_luc’. Drop the unwanted columns except for the three variables mentioned - ‘gdp’, ‘population’ and ‘co2_including_luc’.

year	country	iso_code	population	gdp	cement_co2	cement_co2_per_capita	co2	co2_growth_abs	co2_growth_prct	co2_including_luc
1850	Afghanistan	AFG	3752993.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1851	Afghanistan	AFG	3769828.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1852	Afghanistan	AFG	3787706.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1853	Afghanistan	AFG	3806634.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1854	Afghanistan	AFG	3825655.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Figure. 4.1 Head of the dataset

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 50598 entries, 1850 to 2021
Data columns (total 78 columns):
 #   Column                                      Non-Null Count  Dtype
---  ------                                      --------------  -----
 0   country                                     50598 non-null  object
 1   iso_code                                    41970 non-null  object
 2   population                                  40008 non-null  float64
 3   gdp                                         14564 non-null  float64
 4   cement_co2                                  24974 non-null  float64
 5   cement_co2_per_capita                       22714 non-null  float64
 6   co2                                         31349 non-null  float64
 7   co2_growth_abs                              28944 non-null  float64
 8   co2_growth_prct                             25032 non-null  float64
 9   co2 including luc                           24212 non-null  float64
```

Figure. 4.2 Different columns in the dataset

4.3.1 Identification of missing values

Select the rows pertaining to country ‘India’ in the dataset pertaining to three variables ‘gdp’, ‘population’, ‘co2_including_luc’. The columns with variables ‘gdp’, ‘population’ and ‘co2_including_luc’ need to be checked for missing values. The GDP column has 19% missing data, and the co2_including_luc column has 11% missing data. These missing values can be suitably imputed with mean, median or mode to facilitate further processing and analysis of the data. The density plot has been used to decide the imputation technique to be used for imputing the missing values.

```
country              0.000000
population           0.000000
gdp                 19.186047
co2_including_luc   11.046512
dtype: float64
```

Figure. 4.3 Missing values (percentage) in the dataset

4.3.2 Univariate and Bivariate analysis

Correlations between the variables ‘gdp’, ‘population’ and ‘co2_including_luc’ are checked using the ‘Heat map’ plot. It is found variables ‘co2_including_luc’ and ‘population’ are highly correlated compared to ‘co2_including_luc’ and ’gdp’ variables. The categorical variable present in the dataset is the indexed column ‘country’. For Multiple Linear Regression analysis, the categorical variable is converted to a dummy variable. After the creation of the dummy variable add the results to the original data frame. Drop ‘country’ variable as we have created dummies for it.

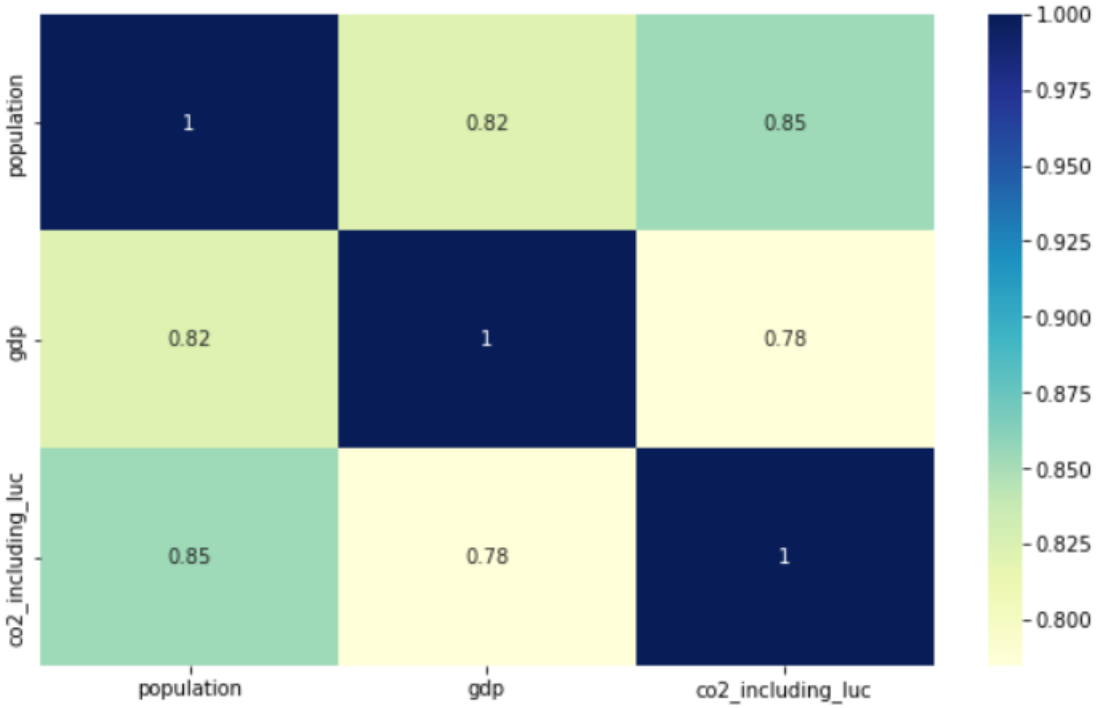


Figure. 4.4 Heat map plot

4.3.3 Treatment of missing values

The numerical columns with variables ‘gdp’, ‘population’ and ‘co2_including_luc’ are plotted using box plot to check for outliers. The distribution of values and how sensitive the outliers are also checked. The GDP column has 19% missing data, and the co2_including_luc column has 11% missing data. The skewness of the data and the density plot will decide the imputation technique to be used. Mean imputation technique is used for missing values of ‘co2_including_luc’ since the missing values are numerical, and the distribution of the variable is approximately normal. Median imputation technique is used for ‘gdp’ since the distribution of values is heavily skewed towards the left of the Inter quartile range.

4.4 Model Building

4.4.1 Feature Scaling

Import the Scikit learn libraries for train - test split of the dataset. Split the dataset into 70% train and 30% test dataset respectively. Scaling variables ‘gdp’, ‘population’ and ‘co2_including_luc’ with the MinMax scaler improves interpretation and accelerates gradient descent convergence. The MinMax scaler uses the data's maximum and minimum values to turn variables into values between zero and one. Apply scaler to all the columns except the dummy variables and formulate a new data frame for training. ‘co2_including_luc’ is the target variable. ‘gdp’ and ‘population’ are independent variables. y_train variable will be ‘co2_including_luc’ and X_train variable will be ‘gdp’ and ‘population’.

4.4.2 Recursive Feature Elimination

Recursive feature elimination (RFE) is a feature selection technique that involves iteratively fitting a model and eliminating the least significant feature(s) until a certain number of features remains. The model's coefficients or feature important attributes are used to rank the features. Through a recursive process of deleting a small number of features per loop, RFE aims to address dependencies and collinearity that may be present in the model.

The Recursive Feature Elimination (RFE) method necessitates a predetermined count of features to retain, however the precise number of acceptable features is frequently unknown beforehand. In order to determine the most suitable number of features, the technique of cross-validation is employed in conjunction with Recursive Feature Elimination (RFE). This approach involves evaluating various subsets of features using RFE, and subsequently selecting the collection of features that yields the highest score.

For building the model, Recursive feature elimination and Linear Regression method is used. Run the RFE with the output number of variables equal to 2. Create X_test data frame with RFE selected variables. Add a constant variable and run the linear model. Check for estimated values of the intercept and the estimated co-efficient. The p-value is a crucial metric in this investigation. The concept of the p-value, as learned in the Statistics course, is employed in the

process of hypothesis testing. In the context of constructing a regression model, it is important to consider the null hypothesis associated with each p-value, which posits that the independent variable in question does not have a significant effect on the dependent variable. The alternative hypothesis posits that the independent variable in question has a significant effect on the response variable. The p-value is a statistical measure that quantifies the likelihood of the null hypothesis being true. Hence, a p-value below the significance level of 0.05 signifies the ability to reject the null hypothesis.

4.4.3 Variance Inflation Factor

The use of examining correlations may be limited, as it is conceivable that a single variable may not fully account for another variable, although a combination of factors may possess the capacity to do so. In order to assess the relationships between variables, the Variance Inflation Factor (VIF) is employed. The Variance Inflation Factor (VIF) is a statistical measure that aids in elucidating the link between one independent variable and all other independent variables. A variable with a Variance Inflation Factor (VIF) greater than 5 is typically considered statistically insignificant. In such cases, the variable's significance, as indicated by its p-value, should be carefully examined, and if necessary, the variable should be eliminated from the analysis.

4.4.4 Residual Analysis

A residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value. When drawing conclusions from a linear regression model, assuming that the residual terms follow a normal distribution is one of the most important assumptions that may be made. As a result, doing an analysis of these residual terms is absolutely necessary before moving on to the next step. Plotting a histogram of the error terms and determining whether or not the error terms are normal is the easiest way for determining whether or not the data is normally distributed.

Residual = Observed Value - Predicted Value

Several key properties of a well-constructed residual plot can be identified as follows:

The data exhibits a pronounced concentration of points in proximity to the origin, accompanied by a notable scarcity of points at greater distances from the origin. The object exhibits symmetry with respect to the origin.

4.4.5 Predictions and Evaluation

Apply scaling to the test dataset as applied earlier to the train dataset. Divide the data frame into y_test which will contain the independent variable 'co2_including_luc' and X_test which will contain the dependent variables 'gdp' and 'population'. Add a constant variable and run the linear model. Make predictions using the predict function. Plot a scatter plot to check the distribution of points y_test vs y_pred. Additionally, an alternative method for assessing the precision of the model was acquired, namely the utilisation of R^2 statistics. The coefficient R^2 quantifies the proportion of variance in the observed data that can be accounted for by the constructed model. The variable consistently assumes a numerical value within the range of zero to one. In broad terms, this concept offers an assessment of the extent to which a model accurately reproduces real-world outcomes. It is determined by the fraction of the overall variation in outcomes that can be accounted for by the model, specifically the anticipated outcomes. In general, a higher value of R^2 indicates a stronger fit between the model and the observed data.

4.5 Vector Auto Regression

For Vector Auto Regression analysis, the initial steps such as from 4.3 Data preparation and exploration to 4.3.3 treatment of missing values remain the same.

The technique of doing exploratory data analysis (EDA) holds significant importance within the realm of statistical modelling, particularly in the construction of a Vector Autoregressive (VAR) model. Exploratory Data Analysis (EDA) entails the comprehensive examination and concise summarization of key data attributes, including distributions, correlations, and trends. This analytical approach aims to get a deeper understanding of the fundamental patterns and interrelationships among variables. Let us employ data visualization techniques. Plot the data to check the level, trend, and seasonality of gdp, population and co2_including_luc variables.

4.5.1 Stationary tests

From a visual standpoint, it is evident that the series exhibits a discernible pattern or seasonal element, which warrants an evaluation of its stationarity. However, it is not possible to definitively assert this claim without conducting tests to verify if the statistical features of the data, such as mean, variance, etc., stay consistent. It is not feasible to provide commentary just by visual inspection of the time series data. There exist explicit exams for the purpose of assessing this.

There exist two official tests. The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test and the Augmented Dickey-Fuller (ADF) test are two often used methods for assessing stationarity in time series data. These tests rely on the principles of hypothesis testing to determine if a given time series is stationary or non-stationary. The unit root test is a widely used method for assessing the stationarity of a time series. A series is considered non-stationary if the presence of a unit root cannot be rejected. Augmented Dickey-Fuller (ADF) test will be used to assess the stationarity of the three variables gdp, population and co2_including_luc. When the p-value is below a certain threshold, typically set at 5%, it indicates that we should reject the null hypothesis. Conversely, if the p-value exceeds the threshold, it suggests that we should not reject the null hypothesis.

4.5.2 Differencing

In order to use auto-regressive models, it was previously said that the time-series data must exhibit stationarity. Therefore, in the case of a non-stationary time series, it is necessary to initially transform it into a stationary time series. There exist two methods for transforming a non-stationary series into a stationary series, which are as follows:

The process of differencing and the concept of transformation refers to the process of altering or changing something from one state or form to another. It involves a fundamental shift. The approach employed to eliminate the presence of a trend and maintain a constant mean in a time series is known as differencing. As implied by its name, the process of differencing involves the calculation of the disparities between successive observations. The process might be likened to taking the difference of a line with a non-zero slope in order to achieve a slope of zero. The process of differencing serves to stabilize the mean of a time series by eliminating or minimising

the effects of variations in the level of the time series. This, in turn, leads to the elimination or reduction of trend and seasonality in the data.

4.5.3 Causality

The formal definition of Granger causality pertains to the examination of whether previous values of variable x contribute to the forecasting of variable yt, given that the impacts of previous values of variable y (and maybe other factors) on yt have already been taken into consideration. If such an occurrence takes place, it is referred to as the "Granger causality" of x on y. The fundamental principle behind VAR is that there exists interdependence among the time series variables within the system. This study aims to assess the null hypothesis that the coefficients of previous values in the regression equation are equal to zero. If the p-value derived from the statistical test is less than the predetermined significance level of 0.05, it is appropriate to reject the null hypothesis. The analysis of Granger causality matrix of the three variables gdp, population and co2_including_luc.

4.6 Building the model

Create a training data set and a testing data set with the data. Import the vector auto regressive model that is located in the library of statistical models. Before we can begin to fit the data to our model, we need to choose the order of the AR(p) function. In order to accomplish this, we will make use of the function model.select_order, which will provide us with the AIC and BIC score along with the number of lags.

4.6.1 Forecasting

The predicted value of the variable GDP and CO2 including land use change (LUC) exhibits a first-order difference, but the variable population demonstrates a second-order difference. In order to achieve similarity with the original data, it is necessary to revert each discrepancy. This process involves incorporating the most recent observations from the original series training data and combining them with a cumulative sum of projected values. The subsequent procedure

involves generating a graphical representation that compares the projected values with the actual values of GDP, population, and CO2 emissions, including land-use change.

4.6.2 Evaluation

The metric in question is the square root of the mean square difference between the predicted values and the actual values within a given dataset. A model's fit to a dataset is considered better when the root mean square error (RMSE) is lower. The provided metric offers an approximation of the model's predictive capability in relation to the goal value, specifically in terms of accuracy. The Root Mean Square Error (RMSE) is a widely accepted metric for quantifying the accuracy of a model's predictions of continuous data. The root mean square error (RMSE) is a suitable metric for estimating the standard deviation σ of a typical observed value based on our model's prediction. This estimation assumes that the observed data can be decomposed into the expected value and a random noise component that is distributed with a mean of zero.

The random noise present in this context encompasses any factors that are not accounted for by our model, such as unidentified variables that may impact the observed data. When the magnitude of the noise, as assessed by the root mean square error (RMSE), is minimal, it typically indicates that our model is proficient in forecasting the observed data. Conversely, if the RMSE is substantial, it generally implies that our model is inadequate in capturing the underlying aspects of the data. The RMSE is assessed for the three variables gdp, population and co2_including_luc.

4.7 Summary

This section explores the implementation of the workflow, encompassing a series of pivotal steps. The initial phase of our study involved doing a thorough examination of the Dataset. This stage was dedicated to understanding the data gathering process and performing preliminary inspections. We also analysed the Dataset's data types and calculated summary statistics to summarize the data. We then did a quality analysis to find and fix missing data during Data Preprocessing. We then focused on Exploratory Data Analysis (EDA) to understand gdp, population, and co2_including_luc.

As a component of exploratory data analysis (EDA), we conducted correlation analysis and generated dummy variables for the categorical variable "country," subsequently incorporating them into the original dataset. In the remaining sections of this chapter, our attention was directed towards the analysis of the data through the utilisation of Multiple Linear Regression employing the Recursive Feature Elimination technique. Furthermore, the evaluation of the results was conducted utilising the R^2 measure. In addition, our study encompassed an examination of Time Series data, including an analysis of its distinct components. Furthermore, we evaluated the stationarity of the data and explored the causality matrix that exists among the three variables. Furthermore, we employed the VAR () function to establish and assess our model by utilising the Root Mean Square Error (RMSE).

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Introduction

Following the comprehensive examination of the methodology and implementation of the established models in the preceding chapters, it is now appropriate to delve into the outcomes and observations derived from the execution of those models. This chapter presents the findings of each iteration utilising Multiple linear regression and Vector Auto Regression models. The evaluation measures, namely R^2 and Root Mean Square Error (RMSE), have been the subject of discussion. The subsequent chapter delves into the handling of missing values and their influence on the model, as well as the concepts of stationarity and causality in relation to the variables' data. The paper also examines the pre-processing techniques employed on the dataset and evaluates their effects. Moreover, the chapter elucidates the process of visualizing the outcomes utilising diverse methodologies.

5.2 Dataset Issues

In order to obtain meaningful results, it is important to address many significant challenges associated with the dataset utilised for research. The primary concern in this context is the multitude of variables present in the dataset and the essential duty of choosing the suitable variables for analysis. In addition, the presence of missing values and the variety of scales pose a substantial problem throughout the data processing phase. Furthermore, the selection of a suitable imputation technique for addressing missing variables has the potential to influence the assessment of the model. To effectively tackle these issues, it is crucial to use thorough data cleansing, imputation, and verification methods to ensure the reliability and representativeness of the results. Moreover, it is crucial to take into account alternative analytical approaches in order to effectively address the limitations inherent in the dataset.

5.3 Univariate analysis results

The box plot analysis of the variable 'gdp' reveals a pronounced left skewness, as depicted in Figure 3.6. As elucidated in the fourth chapter, the variable 'gdp' exhibits missing values that

necessitate imputation. The median imputation technique is employed for the variable 'gdp' due to its significantly left-skewed distribution inside the interquartile range, as seen by the density plot depicted in Figure 3.7. Median imputation is the best method in cases where the distribution of the data is skewed. This is because the median is less affected by outliers compared to the mean.

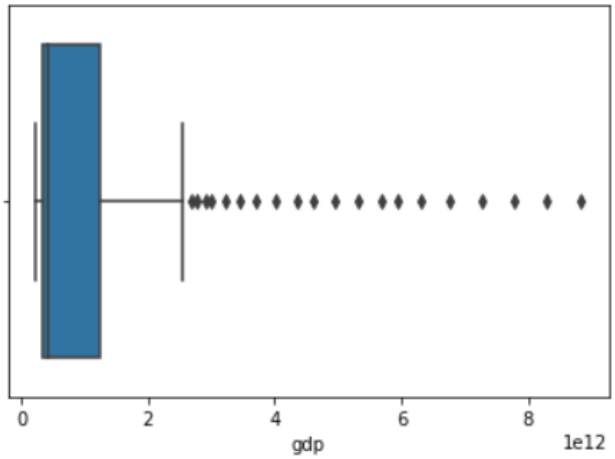


Figure. 5.1 Box plot of gdp

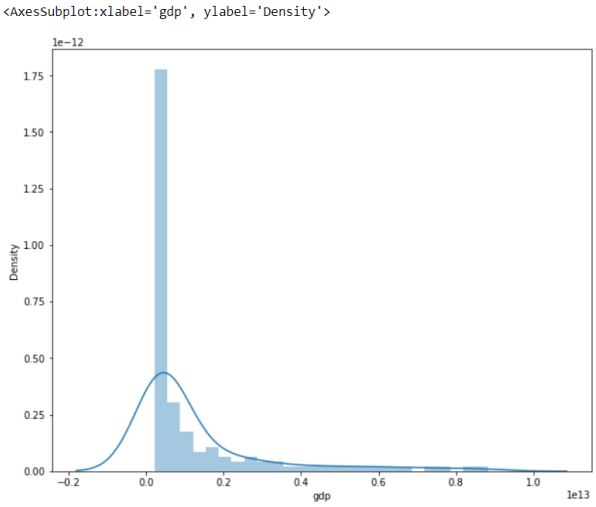


Figure. 5.2 Density plot of gdp

The box plot of the variable 'co2_including_luc' demonstrates that the distribution of the data is relatively normal, as depicted in Figure 3.8. As elucidated in the fourth chapter, the variable 'co2_including_luc' exhibits missing values that necessitate imputation. The mean imputation

technique was employed for the variable 'co2_including_luc' due to its rather normal distribution, as observed in the box plot and density plot depicted in Figure 3.9. Mean imputation is a commonly employed technique for handling missing values in numerical variables, particularly where the distribution of the variable closely approximates a normal distribution.

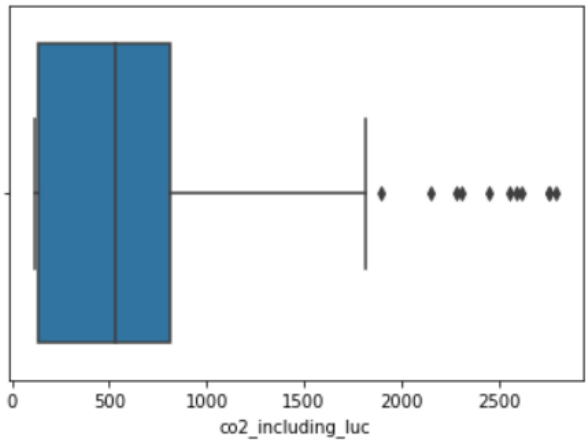


Figure. 5.3 Box plot of co2_including_luc

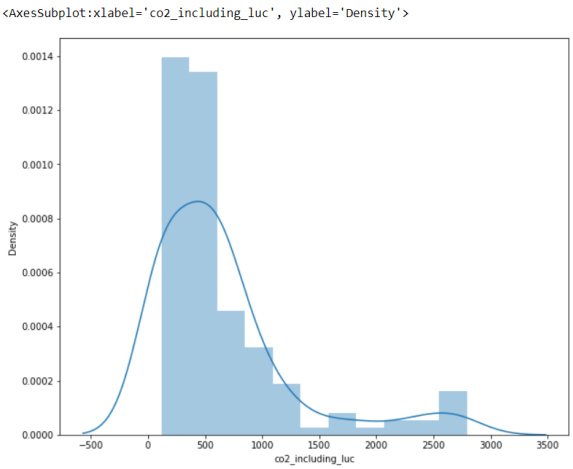


Figure. 5.4 Density plot of co2_including_luc

5.4 Building the model – Multiple Linear Regression

5.4.1 Feature Selection

As described in subsection 5.2 Dataset Issues, the variables 'gdp', 'population', and 'co2_including_luc' exhibit varying scales and ranges of values. In order to tackle this matter, the Min-Max scaler function is utilised to standardize the numbers within the interval of 0 to 1. The variables' values have been normalized, as depicted in Figure 3.10. The normalized values will be utilised for the purposes of constructing the model and assessing its performance.

year	population	gdp	co2_including_luc	India
1945	0.128148	0.026188	0.177763	1
1901	0.049663	0.006702	0.003425	1
1996	0.644075	0.260271	0.339043	1
1942	0.135644	0.025568	0.174988	1
1977	0.358378	0.089860	0.210667	1

Figure. 5.5 Variable values after applying Min - Max Scaler

5.4.2 Recursive Feature Elimination & VIF results

The Recursive Feature Elimination (RFE) method requires a pre-determined number of features to be retained, however the exact number of acceptable features is often unknown in advance. The Recursive Feature Elimination (RFE) technique was employed with a specified output variable count of 2. The OLS regression findings, as depicted in Figure 3.11, reveal that the p-values of all variables are below 0.05. Additionally, Figure 3.12 demonstrates that the Variance Inflation Factor (VIF) values are all less than 5. This finding indicates that the null hypothesis may be rejected, implying that there is no need to eliminate any variables from the model due to the absence of collinearity, thus allowing for further research.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      co2_including_luc   R-squared:                       0.732
Model:                            OLS   Adj. R-squared:                  0.727
Method:                 Least Squares   F-statistic:                     159.7
Date:                Sat, 28 Oct 2023   Prob (F-statistic):           3.60e-34
Time:                        11:02:18   Log-Likelihood:                 93.714
No. Observations:                 120   AIC:                            -181.4
Df Residuals:                     117   BIC:                            -173.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
population     0.5001      0.065      7.701      0.000       0.371       0.629
gdp            0.2666      0.094      2.848      0.005       0.081       0.452
India          0.0565      0.013      4.199      0.000       0.030       0.083
==============================================================================
Omnibus:                       17.261   Durbin-Watson:                   2.375
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               20.259
Skew:                           0.865   Prob(JB):                     3.99e-05
Kurtosis:                       4.029   Cond. No.                         11.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Figure. 5.6 OLS Regression results

	Features	VIF
0	population	3.09
1	gdp	3.09
2	India	1.73

Figure. 5.7 VIF results of variables

5.4.3 Residual analysis

One of the fundamental assumptions underlying linear regression is the normal distribution of error terms. According to the findings shown in Figure 3.13, it can be observed that the error term distribution deviates slightly from the normal distribution. The reason for this is the reduced amount of features that were chosen as part of the Recursive Feature Elimination (RFE) approach. The study was constrained to employing MLR and VAR analysis solely on three

variables, namely 'gdp', 'population', and 'co2_including_luc', due to the specific scope of the research.

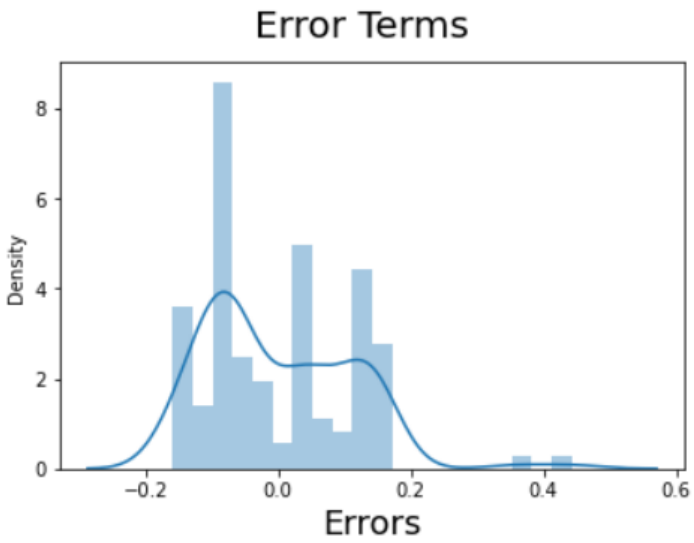


Figure. 5.8 Distribution of Error terms

5.4.4 Model Evaluation

A scatter plot depicting the actual values and expected values has been generated, as illustrated in Figure 5.9 below. A regression plot is a valuable tool for comprehending the linear association between two variables. A regression line is generated based on the given parameters, and then, a scatter plot is constructed using the provided data points. The regression plot of the model is depicted in Figure 5.10 presented below.

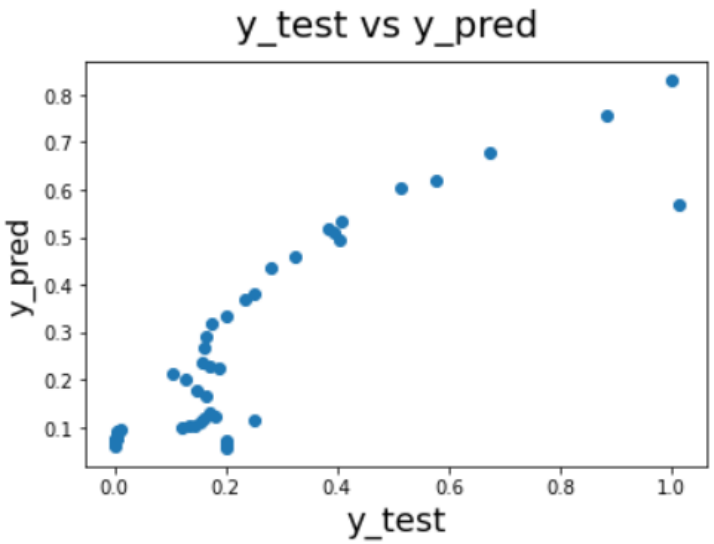


Figure. 5.9 Scatter plot of actual vs predicted values

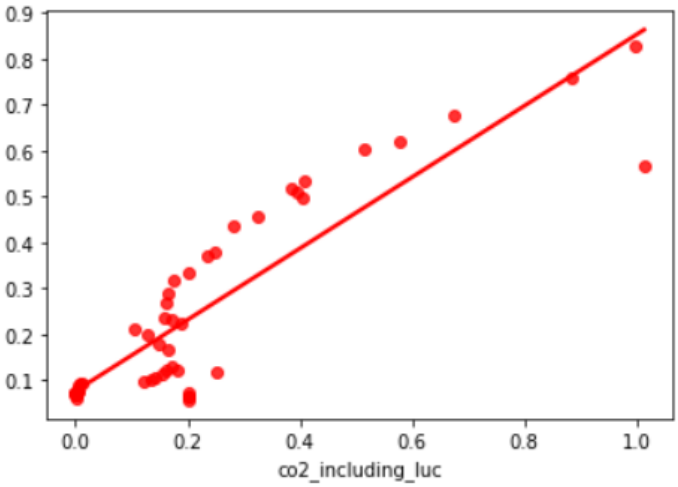


Figure. 5.10 Regression plot of the model

A novel data frame has been created to include actual values, anticipated values, and the corresponding disparities. This will facilitate an assessment of the model's accuracy in predicting the actual values. The disparity between the observed value and the estimated value, as depicted in Figure 5.11, is negligible, hence indicating the model's accuracy. The model's R2 score is 78%, indicating a rather high level of explanatory power.

year	Actual Value	Predicted Value	Difference
1857	0.200176	0.066026	0.134149
1974	0.158461	0.236625	-0.078165
1861	0.001981	0.060991	-0.059010
1865	-0.000170	0.069416	-0.069586
1981	0.164757	0.290430	-0.125673
1909	0.003553	0.091505	-0.087952
1968	0.128986	0.199487	-0.070502
1947	0.180389	0.122034	0.058355
1891	0.005110	0.078227	-0.073117
2000	0.404335	0.495753	-0.091418
1876	0.200176	0.072992	0.127183
1989	0.234201	0.369572	-0.135371
1851	0.200176	0.056959	0.143217
2007	0.513874	0.604398	-0.090524
1951	0.250517	0.116760	0.133757
2015	0.883846	0.757732	0.126114
2002	0.383517	0.519697	-0.136181
1984	0.173489	0.317609	-0.144120
1986	0.200160	0.335991	-0.135831
1935	0.160534	0.120864	0.039670
1925	0.134924	0.101868	0.033056

Figure. 5.11 Difference values of Actual value and predicted value

5.5 Vector Auto Regression

5.5.1 Exploratory data analysis

The utilisation of normalized data has been limited to the application of multiple linear regression analysis. The line graphs depicting the variables ‘gdp’, ‘population’, and ‘co2_including_luc’ have been generated using real data values. The variable ‘population’ has an overall upward trend, with the exception of a brief period of decline in the 1950s. However, the line plots depicted in Figure 5.12 do not demonstrate any discernible pattern or seasonality in GDP or CO2 emissions, including land use change.

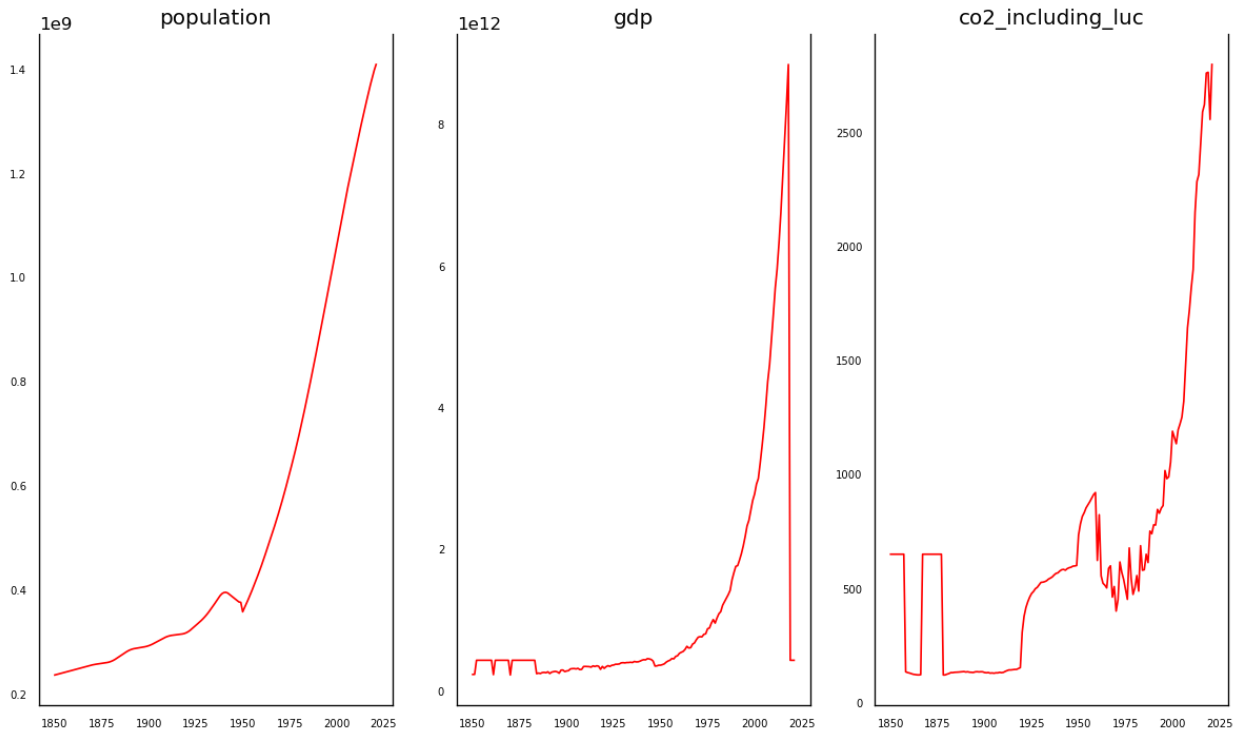


Figure. 5.12 Line plot of gdp, population and co2_including_luc

5.5.2 Stationary test results

The ADF test findings depicted in figures 5.13, 5.14, and 5.15 reveal that the p-values for the variables ‘gdp’, ‘population’, and ‘co2_including_luc’ exceed the threshold of 0.05. Consequently, it can be concluded that the data exhibits non-stationarity. In order to ensure the validity of our analysis, it is necessary for the data pertaining to all three variables to exhibit stationarity, characterised by a consistent mean and variance. The subsequent procedure involves transforming the data into stationary data in order to facilitate subsequent analysis.

```
Augmented Dickey-Fuller Test:
ADF test statistic       -0.203267
p-value                   0.938131
# lags used               3.000000
# observations           168.000000
critical value (1%)      -3.469886
critical value (5%)      -2.878903
critical value (10%)     -2.576027
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

Figure. 5.13 ADF test of population

```
Augmented Dickey-Fuller Test:
ADF test statistic        2.319969
p-value                   0.998968
# lags used               1.000000
# observations           170.000000
critical value (1%)      -3.469413
critical value (5%)      -2.878696
critical value (10%)     -2.575917
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

Figure. 5.14 ADF test of gdp

```
Augmented Dickey-Fuller Test:
ADF test statistic        2.319969
p-value                   0.998968
# lags used               1.000000
# observations           170.000000
critical value (1%)       -3.469413
critical value (5%)       -2.878696
critical value (10%)      -2.575917
Weak evidence against the null hypothesis
Fail to reject the null hypothesis
Data has a unit root and is non-stationary
```

Figure. 5.15 ADF test of co2_including_luc

The variable ‘co2_including_luc’ necessitate a differencing factor of one, while the variables ‘gdp’ and ‘population’ necessitates a differencing factor of two in order to transform the non-stationary data into stationary data, as depicted in figures 5.16, 5.17, and 5.18, respectively. The p-values for all three variables are statistically significant at the 0.05 level. A novel data frame is generated, encompassing all three variables, with differenced data. This data frame is subsequently employed in the VAR analysis.

```
Augmented Dickey-Fuller Test:
ADF test statistic     -1.307838e+01
p-value                 1.893166e-24
# lags used             1.000000e+00
# observations          1.680000e+02
critical value (1%)    -3.469886e+00
critical value (5%)    -2.878903e+00
critical value (10%)   -2.576027e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Figure. 5.16 Differenced ADF test of population

```
Augmented Dickey-Fuller Test:
ADF test statistic       -5.460848
p-value                   0.000003
# lags used               5.000000
# observations          162.000000
critical value (1%)      -3.471374
critical value (5%)      -2.879552
critical value (10%)     -2.576373
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Figure. 5.17 Differenced ADF test of gdp

```
Augmented Dickey-Fuller Test:
ADF test statistic     -6.122516e+00
p-value                 8.785796e-08
# lags used            2.000000e+00
# observations          1.680000e+02
critical value (1%)    -3.469886e+00
critical value (5%)    -2.878903e+00
critical value (10%)   -2.576027e+00
Strong evidence against the null hypothesis
Reject the null hypothesis
Data has no unit root and is stationary
```

Figure. 5.18 Differenced ADF test of co2_including_luc

5.5.3 Causality test results

The response variable (y) is represented by the rows, while the predictors (x) are represented by the columns. If a provided p-value is less than the predetermined significance level of 0.05. Based on the observed value of 0.0474 in the cell located at row 2 and column 3, it is possible to reject the null hypothesis and establish that the variable co2_including_luc_1d_x Granger causes the variable gdp_1d_y as depicted in Figure 5.19.

	population_1d_x	gdp_1d_x	co2_including_luc_1d_x
population_1d_y	1.0000	0.3957	0.3301
gdp_1d_y	0.9046	1.0000	0.0474
co2_including_luc_1d_y	0.0009	0.0000	1.0000

Figure. 5.19 Granger causality matrix

5.5.4 Building the model

The AR (p) value of 15 was chosen to optimise the fit of the data to our model. The model was utilised in our study. The select_order () function is utilised to obtain the values of AIC, BIC, FPE, and HQIC. The Akaike Information Criterion (AIC) exhibits a declining trend as we progressively incorporate a more intricate model. However, beyond a certain threshold, the AIC starts to exhibit an upward trajectory. The reason for this is that AIC penalises these models for their excessive complexity. The VAR (2) model is selected as it identifies the lowest score, followed by an increase in AIC. Therefore, the VAR model of order 2 will be constructed.

VAR Order Selection (* highlights the minimums)

	AIC	BIC	FPE	HQIC
0	88.05	88.11	1.729e+38	88.07
1	87.47	87.72*	9.707e+37	87.57*
2	87.40*	87.84	9.050e+37*	87.58
3	87.45	88.08	9.544e+37	87.71
4	87.48	88.30	9.863e+37	87.81
5	87.57	88.57	1.077e+38	87.98
6	87.47	88.67	9.811e+37	87.96
7	87.52	88.90	1.025e+38	88.08
8	87.59	89.16	1.113e+38	88.23
9	87.52	89.28	1.042e+38	88.24
10	87.58	89.52	1.107e+38	88.37
11	87.50	89.63	1.027e+38	88.36
12	87.50	89.82	1.044e+38	88.45
13	87.56	90.07	1.112e+38	88.58
14	87.63	90.32	1.207e+38	88.72
15	87.57	90.45	1.152e+38	88.74

Figure. 5.20 VAR Order Selection

```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Wed, 15, Nov, 2023
Time:                     20:18:30
--------------------------------------------------------------------
No. of Equations:         3.00000    BIC:                    88.3385
Nobs:                     154.000    HQIC:                   88.0926
Log likelihood:          -7404.73    FPE:                 1.53163e+38
AIC:                      87.9244    Det(Omega_mle):      1.34041e+38
```

Results for equation population_1d

	coefficient	std. error	t-stat	prob
const	172866.993317	183333.849966	0.943	0.346
L1.population_1d	-0.551118	0.080695	-6.830	0.000
L1.gdp_1d	-0.000002	0.000003	-0.667	0.505
L1.co2_including_luc_1d	1901.600590	1897.303276	1.002	0.316
L2.population_1d	-0.188170	0.080411	-2.340	0.019
L2.gdp_1d	-0.000003	0.000003	-0.996	0.319
L2.co2_including_luc_1d	1232.880714	1921.461980	0.642	0.521

Results for equation gdp_1d

	coefficient	std. error	t-stat	prob
const	4426243512.608132	4260831215.263489	1.039	0.299
L1.population_1d	169.826314	1875.422122	0.091	0.928
L1.gdp_1d	-0.864437	0.075955	-11.381	0.000
L1.co2_including_luc_1d	13750995.165877	44094906.786336	0.312	0.755
L2.population_1d	348.982342	1868.825697	0.187	0.852
L2.gdp_1d	-0.398114	0.076522	-5.203	0.000
L2.co2_including_luc_1d	-13170211.081355	44656375.161649	-0.295	0.768

Results for equation co2_including_luc_1d

	coefficient	std. error	t-stat	prob
const	7.206076	7.987454	0.902	0.367
L1.population_1d	-0.000000	0.000004	-0.126	0.900
L1.gdp_1d	-0.000000	0.000000	-0.054	0.957
L1.co2_including_luc_1d	-0.080584	0.082661	-0.975	0.330
L2.population_1d	-0.000000	0.000004	-0.106	0.916
L2.gdp_1d	0.000000	0.000000	0.136	0.892
L2.co2_including_luc_1d	0.059242	0.083714	0.708	0.479

Correlation matrix of residuals

	population_1d	gdp_1d	co2_including_luc_1d
population_1d	1.000000	-0.025195	-0.036200
gdp_1d	-0.025195	1.000000	0.064770
co2_including_luc_1d	-0.036200	0.064770	1.000000

Figure. 5.21 OLS regression results of VAR

5.5.5 Forecasting

It is important to acknowledge that the projected value represents both first-order difference and second-order difference. In order to achieve similarity with the original data, it is necessary to reverse each discrepancy. This process involves incorporating the most recent observations from the training data of the original series and combining them with a cumulative sum of predicted values. The projected values for the variables ‘gdp’, ‘population’, and ‘co2_including_luc’ demonstrate a modest similarity and an ascending pattern when compared to the observed data, as depicted in Figures 5.22, 5.23, and 5.24 correspondingly.

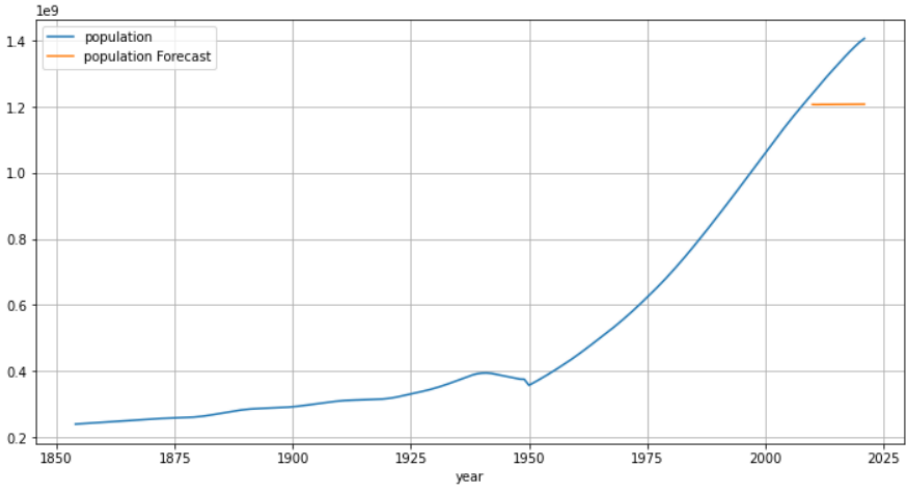


Figure. 5.22 Actual vs Forecast data of population

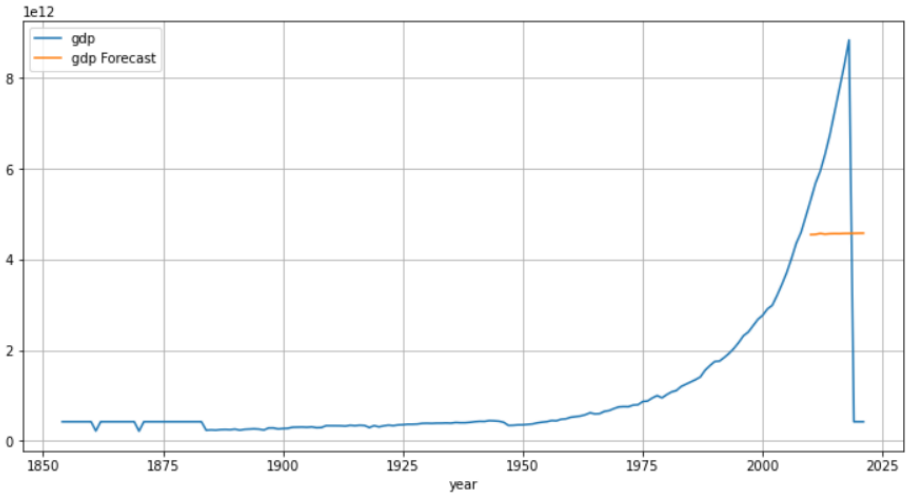


Figure. 5.23 Actual vs Forecast data of gdp

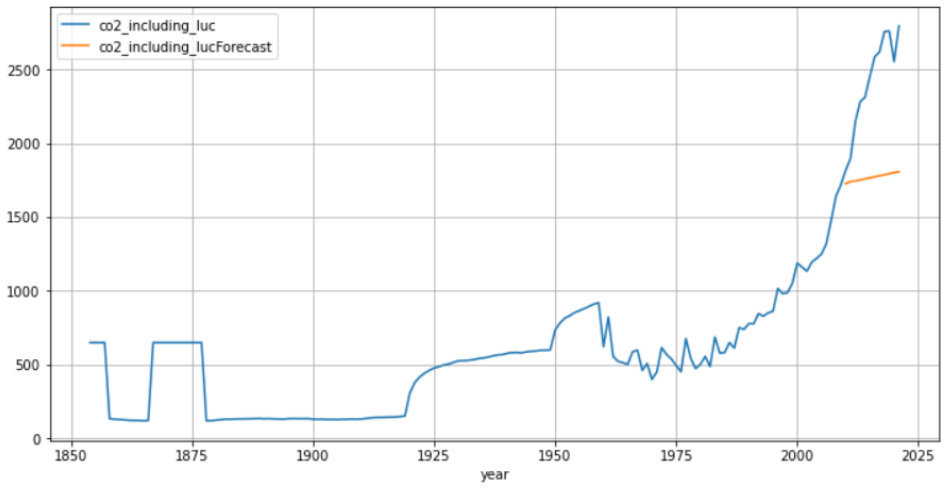


Figure. 5.24 Actual vs Forecast data of co2_including_luc

5.5.6 Evaluation

When the mistakes are squared prior to being averaged, the root mean square error (RMSE) assigns a substantially greater importance to errors of larger magnitude. The root mean square error (RMSE) is most advantageous in situations where the presence of significant errors is highly undesirable. Smaller values are more desirable. The variable ‘gdp’ exhibits values ranging from 2.15E+11 to 8.84E+12. Similarly, the variable ‘population’ showcases values ranging from 0.23 billion to 1.41 billion. Lastly, the variable ‘co2_including_luc’ demonstrates values ranging from 118 million tonnes to 2795 million tonnes. The root mean square error (RMSE) values for ‘gdp’, ‘population’, and ‘co2_including_luc’ are 5.80E+12, 1.2 billion, and 1683 million tonnes, respectively. The RMSE value for the variable ‘co2_including_luc’ falls within the median range, while the RMSE values for ‘population’ and ‘gdp’ are located towards the higher end of the range that encompasses the minimum and maximum values for each respective variable. The results suggest that the reliability of the forecast's accuracy is not completely dependable.

5.5.7 Comparison MLR vs VAR model evaluation

The multiple linear regression (MLR) model has superior accuracy in comparison to the vector autoregression (VAR) model. The multiple linear regression (MLR) model exhibits an R2 value of 78%. Additionally, the variables gdp, population, and co2_including_luc have respective

RMSE values of 5.80E+12, 1.2 billion, and 1683 million tonnes. The RMSE value for the variable ‘co2_including_luc’ falls within the median range, while the RMSE values for ‘population’ and ‘gdp’ are located towards the higher end of the range that encompasses the minimum and maximum values for each respective variable. The projected values for the variables ‘gdp’, ‘population’, and ‘co2_including_luc’ demonstrate a modest similarity and an ascending pattern when compared to the corresponding line graphs of the actual values as shown in figures 5.22, 5.23 and 5.24 respectively.

5.5.8 Summary

Data preprocessing uses mean and median algorithms to find and treat missing data. This study examines the association between gdp, population, and co2_including_luc and how multiple linear regression (MLR) handles categorical variables. Partitioning the data into training and testing sets is the first stage in model construction.

The Recursive Feature Elimination (RFE) approach is employed in Multiple Linear Regression (MLR) analysis, in combination with Variance Inflation Factor (VIF) and p-values. In the context of Vector Autoregression (VAR) analysis, it is essential to assess the properties of stationarity and causality in the data. Applying differencing techniques to the data in achieving stationarity. Utilise the differenced data to construct a model and subsequently fit the data. Revert the initial dataset to its original state to facilitate the process of making predictions. The evaluation of the MLR approach involves the utilisation of the reg plot and the scatter plot to compare the actual data with the projected data. The R^2 measure is employed to evaluate the accuracy of the model. The variable co2_including_luc Granger causes the variable gdp.

The MLR technique achieves an accuracy score of 78%. The VAR technique uses the RMSE measure to evaluate the precision of the model. The root mean squared error (RMSE) values for the variables representing gross domestic product (GDP), population, and carbon dioxide emissions including land use change (co2_including_luc) are 5.80E+12, 1.2 billion, and 1683 million tonnes, respectively. The RMSE value for the variable ‘co2_including_luc’ falls within the median range, while the RMSE values for ‘population’ and ‘gdp’ are located towards the higher end of the range that encompasses the minimum and maximum values for each respective variable.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter provides a comprehensive discussion of the objectives of the thesis, as well as an analysis of the implementation highlights and corresponding results. The proposed solution and outcomes serve to demonstrate the attainment of each objective. Additionally, the chapter delves into the research inquiries and the satisfactory responses obtained via this investigation. Lastly, this study offers several suggestions for future research in this field.

6.2 Discussions and conclusions

As per the research methodology workflow chart, our investigation began with a thorough Dataset analysis. This stage was dedicated to learning the data-gathering process and doing preliminary inspections. We also evaluated the Dataset's data kinds and created summary statistics to summarize the data. Data Preprocessing involves a rigorous quality analysis to find and fix missing values. We then focused on Exploratory Data Analysis (EDA) to understand gdp, population, and co2_including_luc. We performed correlation analysis and created dummy variables for the categorical variable ‘country’, adding them into the original dataset as part of exploratory data analysis (EDA). We focused on data analysis using Multiple Linear Regression and Recursive Feature Elimination. The R^2 metric was used to evaluate the results. Our study also examined Time Series data and its components. We also examined the causation matrix between the three variables and data stationarity. We also used VAR () to build and evaluate our model using Root Mean Square Error. The MLR technique achieves an accuracy score of 78%. The root mean square error (RMSE) values for the variables representing gross domestic product (GDP), population, and carbon dioxide emissions including land use change (CO2_including_luc) are 5.80E+12, 1.2 billion, and 1683 million tonnes, respectively. The RMSE value for the variable ‘co2_including_luc’ falls within the median range, while the RMSE values for ‘population’ and ‘gdp’ are located towards the higher end of the range that encompasses the minimum and maximum values for each respective variable. The multiple linear regression (MLR) model has superior accuracy in comparison to the vector autoregression (VAR) model.

Let's revisit the research questions and try to address the initial queries:

- How carbon dioxide (CO2) emissions, population size, and economic growth are correlated?
  - The 'Heat map' plot compares 'gdp', 'population', and 'co2_including_luc'. 'co2_including_luc' and 'population' are highly correlated compared to 'gdp' and 'co2_including_luc'.
- Which of the factors, population, or economic growth, exhibits a causal relationship with CO2 emissions?
  - Based on the obtained value of 0.0474 and establish that the variable co2_including_luc_ Granger causes the variable gdp. Hence, there exists a significant relationship between gdp and co2_including_luc.
- Which model MLR or VAR gives highest performance of accuracy in terms of evaluation?
  - The MLR model has the highest accuracy (78% R^2) compared to VAR model RMSE values of gdp, population, which are in the more extreme ranges of each variable while RMSE value of co2_including_luc is in the median range.

Let us reexamine the study's objectives to determine if they have been successfully achieved.

- To analyze and identify the relationship between economic growth and environmental degradation.
  - 'co2_including_luc' and 'population' are highly correlated compared to 'gdp' and 'co2_including_luc'.
- To explore the causality between economic growth and environmental degradation.
  - Based on the 0.0474 value, it can be concluded that the variable co2_including_luc_ Granger causes gdp. Thus, gdp and co2_including_luc are significantly related.
- To assess the VAR & Multiple Linear regression models and identify the most precise one to find the relationship between economic growth and environmental degradation.
  - The MLR model has superior accuracy, as evidenced by its high R^2 value of 78%. In contrast, the VAR model's RMSE values of gdp, population, which are in the more extreme ranges of each variable while RMSE value of co2_including_luc is in the median range. The accuracy percentage of R^2 value is comparatively lower because of a smaller number of attributes chosen for the study.

- To use the Vector Autoregressive Model of Timeseries Modeling to forecast CO2 levels of actual values vs predicted values from 2011 to 2021 based on historical data.
    - The forecasted values for the variables gdp', 'population', and 'co2_including_luc' demonstrate a significant similarity to the observed data.

6.3 Future Recommendations

- The process of eliminating outliers from the GDP variable will be conducted, followed by an analysis of the dataset using the Multiple Linear Regression (MLR) and Vector Autoregression (VAR) models.
- The present investigation aims to examine the utilisation of Multiple Linear Regression (MLR) and Vector Autoregression (VAR) models by employing a dataset spanning a period of 50 years, from 1960 to 2021. This approach is adopted in order to mitigate issues related to data redundancy and the handling of outliers, which may arise when employing historical data from as far back as 1850 to 2021.
- The incorporation of supplementary variables, such as forest and green cover, as well as the consideration of CO2 emissions stemming from electricity production and the transportation sector, which are recognised as the primary sources of CO2 emissions in India. Coal constitutes around 60% of the predominant energy source utilised for electrical generation in India. These factors have the potential to impact CO2 emissions and can also contribute to the formulation of policy decisions.
- The investigation of additional greenhouse gases, namely methane (CH4), nitrous oxide (N2O), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), Sulphur hexafluoride (SF6), and nitrogen trifluoride (NF3), aims to examine their causative relationship and linkage with pollution and the economy.
- Identifying cities that are major CO2 and GHG sources. These elements will help formulate economic and regulatory modifications and implement carbon dioxide emission mitigation measures.

References

Abbasi, K.R., Kirikkaleli, D. and Altuntaş, M., (2022) Carbon dioxide intensity of GDP and environmental degradation in an emerging country. *Environmental Science and Pollution Research*, 2956, pp.84451–84459.

Akorede, Y.F. and Afroz, R., (2020) The relationship between urbanization, CO2 emissions, economic growth and energy consumption in Nigeria. *International Journal of Energy Economics and Policy*, 106, pp.491–501.

Alhindawi, R., Nahleh, Y.A., Kumar, A. and Shiwakoti, N., (2020) Projection of greenhouse gas emissions for the road transport sector based on multivariate regression and the double exponential smoothing model. *Sustainability (Switzerland)*, 1221, pp.1–18.

Aminata, J., Nugroho, S.B.M., Atmanti, H.D., Agustin, E.S.A.S., Wibowo, A. and Smida, A., (2022) Economic Growth, Population, and Policy Strategies: Its Effecton CO2 Emissions. *International Journal of Energy Economics and Policy*, 124, pp.67–71.

Ardakani, M.K. and Seyedaliakbar, S.M., (2019) Impact of energy consumption and economic growth on CO2 emission using multivariate regression. *Energy Strategy Reviews*, 26.

Aslam, B., Hu, J., Shahab, S., Ahmad, A., Saleem, M., Shah, S.S.A., Javed, M.S., Aslam, M.K., Hussain, S. and Hassan, M., (2021) The nexus of industrialization, GDP per capita and CO2 emission in China. *Environmental Technology and Innovation*, 23.

Das, N., Gangopadhyay, P., Bera, P. and Hossain, M.E., (2023) Investigating the nexus between carbonization and industrialization under Kaya's identity: findings from novel multivariate quantile on quantile regression approach. *Environmental Science and Pollution Research*, 3016, pp.45796–45814.

Garidzirai, R., (2020) Time series analysis of carbon dioxide emission, population, carbon tax and energy use in South Africa. *International Journal of Energy Economics and Policy*, 105, pp.353–360.

Haldar, S. and Sharma, G., (2022) Impact of urbanization on per capita energy use and emissions in India. *International Journal of Energy Sector Management*, 161, pp.191–207.

Hao, Y., (2022) The relationship between renewable energy consumption, carbon emissions, output, and export in industrial and agricultural sectors: evidence from China. *Environmental Science and Pollution Research*, 2942, pp.63081–63098.

Hosseini, S.M., Saifoddin, A., Shirmohammadi, R. and Aslani, A., (2019) Forecasting of CO2 emissions in Iran based on time series and regression analysis. *Energy Reports*, 5, pp.619–631.

Kirikkaleli, D., (2020) New insights into an old issue: exploring the nexus between economic growth and CO 2 emissions in China. [online] Available at: https://doi.org/10.1007/s11356-020-10090-x.

Kumari, S. and Singh, S.K., (2022) Machine learning-based time series models for effective CO2 emission prediction in India. *Environmental Science and Pollution Research*.

Kutlu, G. and Örün, E., (2022) The effect of carbon dioxide emission, GDP per capita and urban population on health expenditure in OECD countries: a panel ARDL approach. *International Journal of Environmental Health Research*.

Li, Y., (2023) An Empirical Study on the Factors Influencing Carbon Emissions in Heilongjiang Province Based on VAR Model. In: *E3S Web of Conferences*. EDP Sciences.

Lin, B. and Xu, B., (2020) How does fossil energy abundance affect China's economic growth and CO2 emissions? *Science of the Total Environment*, 719.

Meirun, T., Leonardus, &, Mihardjo, W.W., Haseeb, M., Abdul, S., Khan, R. and Jermsittiparsert, K., (2020) The dynamics effect of green technology innovation on economic growth and CO 2 emission in Singapore: new evidence from bootstrap ARDL approach. [online] Available at: https://doi.org/10.1007/s11356-020-10760-w.

Morelli, G. and Mele, M., (2020) Energy consumption, CO2 and economic growth nexus in Vietnam. *International Journal of Energy Economics and Policy*, 102, pp.443–449.

Mosikari, T.J. and Eita, J.H., (2020) CO2 emissions, urban population, energy consumption and economic growth in selected African countries: A Panel Smooth Transition Regression (PSTR). *OPEC Energy Review*, 443, pp.319–333.

Ngong, C.A., Bih, D., Onyejiaku, C. and Onwumere, J.U.J., (2022) Urbanization and carbon dioxide (CO2) emission nexus in the CEMAC countries. *Management of Environmental Quality: An International Journal*, 333, pp.657–673.

Pachiyappan, D., Ansari, Y., Alam, M.S., Thoudam, P., Alagirisamy, K. and Manigandan, P., (2021) Short and long-run causal effects of co2 emissions, energy use, gdp and population growth: Evidence from india using the ardl and vecm approaches. *Energies*, 1424.

Qayyum, M., Yu, Y., Nizamani, M.M., Raza, S., Ali, M. and Li, S., (2022) Financial Instability and CO2 Emissions in India: Evidence from ARDL Bound Testing Approach. *Energy and Environment*.

Rahman, M.M., Saidi, K. and Mbarek, M. Ben, (2020) Economic growth in South Asia: the role of CO2 emissions, population density and trade openness. *Heliyon*, 65.

Salari, M., Javid, R.J. and Noghanibehambari, H., (2021) The nexus between CO2 emissions, energy consumption, and economic growth in the U.S. *Economic Analysis and Policy*, 69, pp.182–194.

Shaari, M.S., Abidin, N.Z., Ridzuan, A.R. and Meo, M.S., (2021) The impacts of rural population growth, energy use and economic growth on co2 emissions. *International Journal of Energy Economics and Policy*, 115, pp.553–561.

Shabani, E., Hayati, B., Pishbahar, E., Ghorbani, M.A. and Ghahremanzadeh, M., (2021) A novel approach to predict CO2 emission in the agriculture sector of Iran based on Inclusive Multiple Model. *Journal of Cleaner Production*, 279.

Sikder, M., Wang, C., Yao, X., Huai, X., Wu, L., KwameYeboah, F., Wood, J., Zhao, Y. and Dou, X., (2022) The integrated impact of GDP growth, industrialization, energy use, and urbanization on CO2 emissions in developing countries: Evidence from the panel ARDL approach. *Science of the Total Environment*, 837.

Singh, K. and Upadhyaya, S., (2012) *Outlier Detection: Applications And Techniques*. [online] Available at: www.IJCSI.org.

Tran, Q.H., (2022) The impact of green finance, economic growth and energy usage on CO2 emission in Vietnam – a multivariate time series analysis. *China Finance Review International*, 122, pp.280–296.

Uzair Ali, M., Gong, Z., Ali, M.U., Asmi, F. and Muhammad, R., (2022) CO2 emission, economic development, fossil fuel consumption and population density in India, Pakistan and Bangladesh: A panel investigation. *International Journal of Finance and Economics*, 271, pp.18–31.

Wang, Q. and Li, L., (2021) The effects of population aging, life expectancy, unemployment rate, population density, per capita GDP, urbanization on per capita carbon emissions. *Sustainable Production and Consumption*, 28, pp.760–774.

Wang, S., Zeng, J. and Liu, X., (2019) Examining the multiple impacts of technological progress on CO2 emissions in China: A panel quantile regression approach. *Renewable and Sustainable Energy Reviews*, 103, pp.140–150.

Yang, H., Lu, Z., Shi, X., Muhammad, S. and Cao, Y., (2021) How well has economic strategy changed CO2 emissions? Evidence from China's largest emission province. *Science of the Total Environment*, 774.

Yunita, R., Gunarto, T., Marselina, M. and Yuliawan, D., (2023) The Influence of GDP per Capita, Income Inequality, and Population on CO2 Emission (Environmental Kuznet Curve Analysis in Indonesia). *International Journal of Social Science, Education, Communication and Economics (SINOMICS JOURNAL)*, [online] 22, pp.217–230. Available at: https://sinomicsjournal.com/index.php/SJ/article/view/130.

Ritchie, H., (2020) *CO_2 and Greenhouse Gas Emissions*. [online] Our World in Data. Available at: https://ourworldindata.org/co2-emissions.

Anon (2023) *Linear Regression: Introduction to Multiple Regression - codingstreets*. [online] Linear Regression: Introduction to Multiple Regression - codingstreets. Available at: https://codingstreets.com/linear-regression-introduction-to-multiple-regression/ [Accessed 19 Aug. 2023].

Washington, DC :World Resources Institute, (2022) | *Greenhouse Gas (GHG) Emissions | Climate Watch*. [online] | Greenhouse Gas (GHG) Emissions | Climate Watch. Available at: https://www.climatewatchdata.org/ghg-emissions?chartType=percentage&end_year=2020&gases=co2&start_year=1990 [Accessed 13 Aug. 2023].

APPENDIX A: RESEARCH PROPOSAL

A STUDY OF THE RELATIONSHIP BETWEEN GDP, POPULATION & CO2 EMISSIONS BASED ON TIME SERIES

CHETAN R TIPPA

Research Proposal

AUGUST 2023

Abstract

China, US & India are the largest emitters of CO2 emissions in the world. Motivated by the commitment of India at COP26, we study the relationship between GDP, population & CO2 emissions of India. This study aims to investigate the link between economic development and environmental deterioration. The study uses the "Our World in Data" dataset on India's GDP, population, and CO2 emissions including changes in land use from 1850 to 2021. We employ both the VAR model & Multiple linear regression model. The study's scope is restricted to comparing the VAR model with the Multiple linear regression model, to identify the most accurate model for the study through model evaluation. The Granger causality test is conducted to ascertain the relationship between variables prior to constructing the model. By performing Granger's causality test both unidirectional and bidirectional causality can be found between the variables. The VAR model is evaluated using accuracy metrics RMSE & Multiple linear regression using R^2. A projection of forthcoming patterns in carbon dioxide (CO2) emissions is performed for the ten-year period from 2021 to 2031, utilising historical data spanning from 1850 to 2021. The findings derived from the study will provide valuable insights for the government to inform and guide potential policy modifications.

Keywords: India, GDP, Population, CO2 emissions, VAR, Multiple Linear regression, RMSE, R^2, Granger's causality test.

LIST OF FIGURES

Figure 1. GHG emissions country-wise percentage from 1990 - 2020 6
Figure 2. CO2 emissions country-wise percentage from 1990 - 2020 6
Figure 3. VAR & Multiple linear regression Framework illustrating workflow..... 13
Figure 4. Gantt chart – Project plan and execution 20

LIST OF ABBREVIATIONS

LF-VAR…….. Long Frequency Vector Auto Regression
OLS…..... Ordinary Least Squares
CCR…… Correct classification rate
GDP…… Gross Domestic Product
COP…… Conference of Parties
luc…… Land use change
SDG…… Sustainable Development Goal

Table of Contents

Abstract ii

LIST OF FIGURES iii

LIST OF ABBREVIATIONS iii

1. Background 2

2. Related Research 5

3. Research Questions (If any) 9

4. Aim and Objectives 10

5. Significance of the Study 10

6. Scope of the Study 11

7. Research Methodology 11

7.1 Dataset Description 12

7.2 Exploratory data analysis 13

7.3 Vector Auto Regression (VAR) 13

7.4 Check for Stationarity 13

7.5 Granger's Causality Test 14

7.6 Split & Model the Data 14

7.7 Fit the data using the best lag value 14

7.8 Forecast, predict & plot the results 14

7.9 Multiple linear regression 15

7.10 Spilt the data into the train & test set 15

7.11 Building the model 16

7.12 Residual analysis 16

7.13 Making Predictions 16

7.14 Model evaluation 17

7.15 Interpretation of coefficients 17

8. Requirements Resources 17

9. Research Plan 18

References 19

1. Background

Climate change is mostly driven by the emissions of greenhouse gases (GHGs) that are produced by human activities. Approximately 60% of greenhouse gas (GHG) emissions are attributed to a mere 10 countries, whilst the 100 countries with the lowest emissions collectively account for less than 3% of total GHG emissions. Energy accounts for around 75% of world emissions, with agriculture being the subsequent major contributor. In the energy sector, the primary source of emissions is the electricity and heat generation sector, which is subsequently followed by the transportation and manufacturing sectors. The sector of land use, land use change and forestry (LULUCF) have a dual role as both a source and sink of emissions, making it a crucial component in achieving net-zero emissions. In the year 2020, the combined contributions of China and the United States accounted for 40% of the total GHG emissions on a worldwide scale. Subsequently, the European Union, India, the Russian Federation, and Indonesia followed suit in terms of their respective emissions (Washington, DC: World Resources Institute, 2022).

In the year 2020, China emerged as the primary source of carbon dioxide (CO2) emissions, accounting for around 30% of the global total. The United States followed closely behind, contributing approximately 12% of the global CO2 emissions, while India accounted for approximately 6.2% of the global emissions (Washington, DC: World Resources Institute, 2022). The Prime Minister of India, Narendra Modi, has expressed his dedication to attaining carbon neutrality by the year 2070. Concurrently, leaders from around the world have reached a collective consensus to restrict the increase in global average temperature to a level below 2°C. This commitment was established during the 26th Conference of the Parties (COP 26) convened in Glasgow.

The Earth's greenhouse gas emissions trap solar heat. This causes climate change and global warming. The earth is warming faster than ever. Warmer temperatures are altering weather patterns and nature's balance. This endangers us and all life on Earth.

Climate change increases heatwaves, strong storms, drought, and species extinction. Rising temperatures have caused more storms, deaths, and economic losses. Tropical cyclones, hurricanes, and typhoons are fuelled by the warming ocean. Carbon dioxide absorption by the ocean threatens coral reefs and marine life. Climate change might wipe out one million terrestrial and marine species in the next few decades.

Fish, crops, and animals die or produce less due to climate change, causing world hunger. The ocean acidifies, reducing crops and livestock health. Climate change is humanity's biggest health risk, killing 13 million people annually and straining healthcare systems.

Floods destroy urban slums, make outdoor labour tougher, and reduce crop yields due to climate change. The most vulnerable and unprepared countries for climate change send the most refugees.

The study aims to examine the correlation between electricity consumption and real gross domestic product (GDP), with a specific emphasis on their impact on environmental deterioration, specifically in terms of carbon dioxide (CO2) emissions. The study employed annual time series data spanning from 1971 to 2017 to investigate the presence of causal linkages, both in the short and long term, using the Dickey-Fuller test, Johansen cointegration analysis, and Granger causality analysis. The research revealed a sustained correlation between energy consumption, economic growth, and carbon dioxide (CO2) emissions in the short term (Pandey and Rastogi,2019).

The impact of Indian CO2 emissions on economic growth, using the SDGs framework and 2030 targets. It examines GDP, energy intensity, and CO2 emissions, finding a one-way link between energy use and GDP. The research suggests that conservative energy measures may hinder economic growth due to energy dependence. The study suggests India should switch to renewable energy for cleaner energy and eco-friendly ecosystems, given global environmental awareness (Udemba et al., 2021).

Figure. 1 GHG emissions country-wise percentage from 1990 – 2020 (Washington, DC: World Resources Institute, 2022)

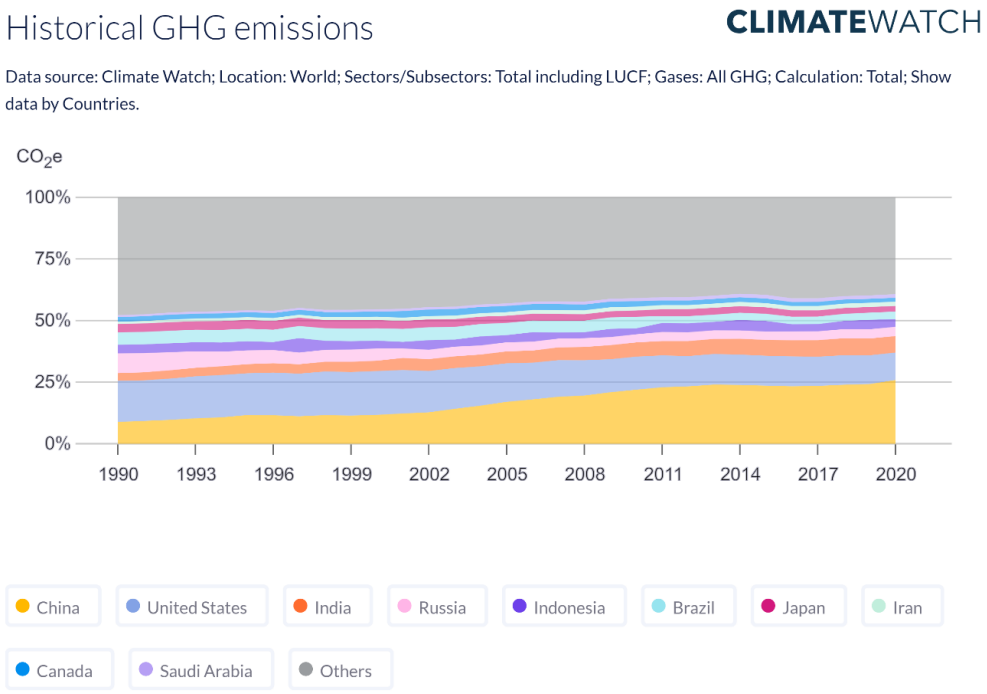
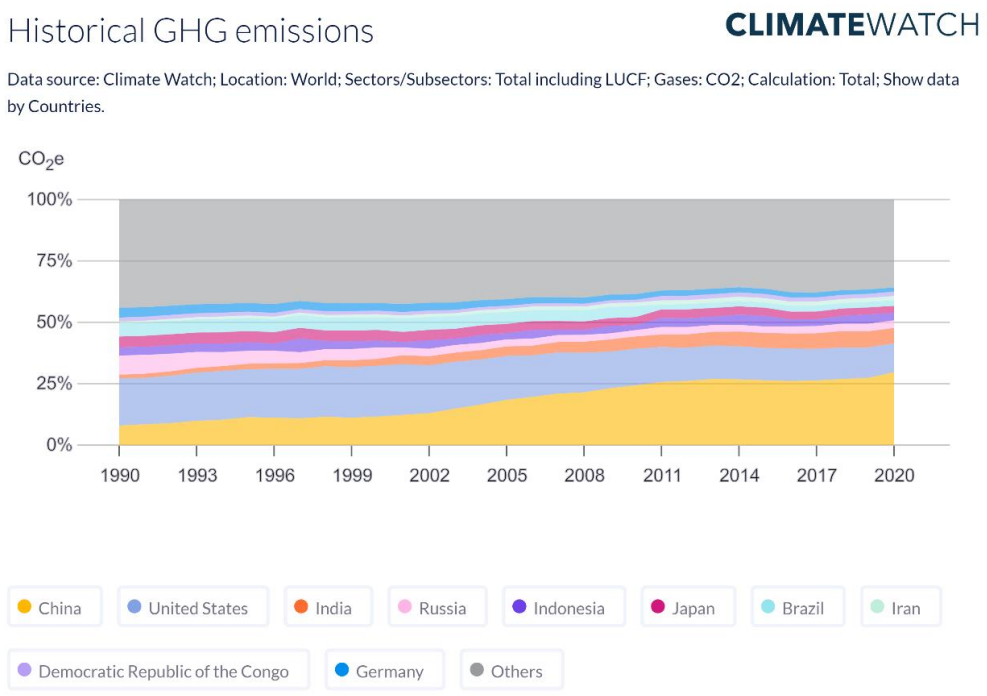


Figure. 2 CO2 emissions country-wise percentage from 1990 – 2020 (Washington, DC: World Resources Institute, 2022)



2. Related Research

The association between electricity use and real GDP is examined in this empirical study. This relationship's effects on environmental degradation, particularly CO2 emissions, are also examined. This study examines the links between real GDP, electricity usage, and CO2 emissions. The investigation used 1971–2017 annual time series data. The Dickey-Fuller test determines variable stationarity. Johansen cointegration and Granger causality are used to examine the short- and long-term causal relationships between power consumption, real GDP, and CO2 emissions. Cointegration among specific combinations of the two variables is shown by the Johansen cointegration test, demonstrating a long-term link between the defined variables. Energy use also causes economic development and CO2 emissions in the near run. (Pandey and Rastogi,2019).

Long-term relationships between health expenditure, CO2 emissions, and GDP per capita in 18 OECD countries were examined using annual time-series data from 1975 to 2017 using McNown et al.'s bootstrap autoregressive-distributed lag (ARDL) cointegration model. The Netherlands, New Zealand, and the US have cointegration in real GDP per capita, health expenditure, and CO2 emissions. The major results reveal a short-run relationship between the three variables. Health spending and GDP growth in Germany and the US, CO2 emissions and GDP growth in Canada, Germany, and the US, and health expenditure and CO2 emissions in New Zealand and Norway are bidirectionally correlated. The results reveal unidirectional causality in other nations (Wang et al., 2019).

Carbon dioxide emissions were examined in relation to energy use and economic growth. Regression, pooled OLS regression, fixed effects, Granger causality, and panel cointegration tests were used in the study. Data from 70 nations from 1994 to 2013 was analysed. The study found a reciprocal causal relationship between the population, capital stock, economic growth, and CO2 emissions, but not energy consumption. Cointegration tests show economic growth, CO2 emissions and energy use are linked. The study stressed the need for a global transition to mitigate carbon emissions and prioritise climate financing, including the mobilisation of public, private, and alternative financial resources to invest in renewable energy and environmentally sustainable projects. (Osobajo et al., 2020).

Environmental deterioration in developing countries is caused by using non-renewable energy for economic growth. This study uses 1965–2015 yearly time series data to examine Pakistan's economic growth, energy consumption, and CO2 emissions. The ARDL estimations show that economic growth and energy consumption increase CO2 emissions in Pakistan, both short-term and long-term. These findings suggest that Pakistani governments should encourage and utilise renewable energy sources to address rising energy demand. Replace coal, gas, and oil with renewable energy to cut CO2 emissions and sustain Pakistan's economy (Khan et al., 2020).

Indian CO2 emissions, a major polluter, harm the ecology despite the expansion. Indian and other CO2 emissions and economic growth literature is divided. The SDGs framework and 2030 targets are used to study India's energy, climate, and economic growth. The study investigates Indian GDP, energy intensity, and CO2 emissions. Trade openness and energy use reduce omitted variable bias in a carbon income function. The Autoregressive Distributed Lag (ARDL) method and modified Wald test of Toda Yamamoto (T-Y) are used for 1975–2017 yearly time series data. Long-term equilibrium exists for the variables. CO2 emissions statistically damage trade and growth. Energy-induced economic expansion supports ARDL regression. Energy use and GDP have a one-way link, suggesting that conservative energy measures will hurt economic growth owing to energy dependence. Given global environmental awareness, India should switch to renewable energy for cleaner energy and eco-friendly ecosystems (Udemba et al., 2021).

State-level CO2 emissions, energy consumption, and GDP from 1997 to 2016 are examined. GDP and energy consumption—total, non-renewable, renewable, industrial, and residential—affect CO2 emissions among states, according to static and dynamic models. Both static and dynamic models reveal a long-term relationship between state-level CO2 emissions and energy use. Non-renewable, industrial, and residential energy use raises CO2 emissions, while renewable energy decreases them. CO2 emissions and GDP have an inverted-U relationship, supporting the Environmental Kuznets Curve (EKC) hypothesis across states. These results are consistent and robust across states using static and dynamic models. This can assist policymakers reduce CO2 emissions in all U.S. states. (Salari et al., 2021).

Urbanisation, energy use, gross capital formation, CO2 emissions, and economic growth in South Korea are examined considering the UN Sustainable Development Goals (SDGs), which emphasise energy access (SDG-7) and sustainable development (SDG-8). This link has not been studied using advanced econometrics. The study uses 1965–2019 data. This association is studied using ARDL, DOLS, and completely modified OLS. Gradual shift and wavelet coherence dictate causation. The ARDL bounds test links variables of interest long-term. CO2 emissions significantly harm economic growth, emphasising the need for South Korea to switch to renewable energy to promote sustainable energy and a sustainable ecosystem. Conservative energy policies may restrict economic growth, supporting the energy-induced growth hypothesis. A one-way causation between energy use and GDP suggests South Korea should avoid stringent energy laws that could limit economic growth. This significantly impacts South Korean GDP and macroeconomic indices (Adebayo et al., 2021).

The growth of Vietnam's per capita income from 1990 to 2019 was evaluated in relation to the use of non-renewable energy, renewable energy, and CO2 emissions. The study cointegrated yearly Vietnam data using Autoregressive Distributed Lag (ARDL). Long-term, non-renewable energy use raises per capita income, whereas CO2 emissions lower it. Non-renewable and renewable energy use changes enhance Vietnam's per capita income. Prior non-renewable energy usage changes hampered Vietnamese income growth. This study illuminates the growth effects of renewable, non-renewable, and CO2 emissions. The findings help Vietnam establish a long-term economic growth strategy. (Nguyen and Le, 2022).

Economic growth, industrial production, and energy consumption determine per capita carbon dioxide emissions, which the proposed study evaluates. Countries with high worldwide CO2 emissions per person are the focus of this study. Data were analysed using panel regression with heterogeneous time trends. After extensive examination, a non-effect panel regression model with heterogeneous time trends is best for our study. (Puntoon et al.,2022) found a positive link between energy use and CO2 emissions. However, economic development and industrial production have a weaker link with CO2 emissions. This study examines how economic development, industrial production, and energy consumption affect CO2 emissions per capita. The study analyses nations with the highest per-capita CO2 emissions using panel regression with heterogeneous time trends. After careful consideration, a non-effect panel regression with heterogeneous time trends is best. Energy consumption has a strong positive

effect on CO2 emissions, while economic expansion and industrial production have a weaker effect.

India's economic growth, energy consumption, foreign direct investment, carbon dioxide emissions, population density, inflation, and agricultural land are evaluated. Autoregressive distributed lag models employ 1985–2019 annual time-series data. Fully modified ordinary least squares, dynamic OLS, canonical co-integrating regression, variance decomposition, and impulse response function demonstrate the model's robustness and uniqueness. Variable cointegration suggests a long-term relationship. Energy use, CO2 emissions, inflation, and agricultural land affect short-term economic growth. Carbon emissions boost long-term economic growth, but inflation and agricultural land damage it. The economic expansion causes energy consumption, carbon emissions, and agricultural land, according to Granger causality. The research improves policymakers' economic growth indicator knowledge and enriches the literature. To promote economic growth and minimise carbon emissions, policymakers should set renewable energy consumption goals. An eco-friendly and economically sustainable future is promoted. (Singh and Kaur, 2022).

China's 1990–2020 renewable energy consumption, output, export, and CO2 emissions are examined. Econometrics supports feedback theory with a two-way causal link. In the medium term, industrial and agricultural export and renewable energy consumption are negatively correlated, supporting growth theory. This means peak export demand may require additional fossil fuels due to renewable energy supply constraints. The findings back China's long-term pollution control and renewable energy objectives. (Hao, 2022).

The relationship between GDP CO2 intensity and environmental deterioration in the rising nation Turkey from 1990 to 2018 is examined. Economic development, foreign direct investment, and renewable energy usage are controlled to investigate their consequences. The bounds test fully modified ordinary least squares (FMOLS), Gregory and Hansen cointegration test, dynamic OLS, nonlinear autoregressive distributed lag (NARDL) model, and CCR are used to evaluate these factors' effects on environmental degradation in Turkey. Environmental degradation in Turkey is linked to GDP CO2 intensity, and lowering GDP intensity lessens it, according to empirical studies. Economic growth also promotes environmental sustainability. These findings suggest that aggressive government measures can tackle environmental issues (Abbasi et al., 2022).

In Taiwan, the MF-VAR model is applied for the first time. Using a high-frequency dataset, economic development and carbon dioxide emissions are investigated from 1970 to 2019. Energy use is primarily employed as a control variable to mitigate the influence of confounding variables that may be absent from the study. Based on forecast error variance decomposition and the Granger causality test, the empirical results suggest that the MF-VAR model shows variable relationships better than the classic VAR model. The empirical LF-VAR model demonstrates a positive relationship between primary energy consumption and economic growth, leading to subsequent increases in CO2 emissions. The MF-VAR model in Taiwan reveals a bidirectional and causal association among economic progress, carbon dioxide emissions, and primary energy use. The findings of this analysis indicate that the implementation of a robust energy plan is important for Taiwan to enhance its economic growth (Chang et al., 2023).

G7 CO2 emissions and economic growth are examined from 1820 to 2021. Emissions and economic growth asymmetry are examined using quantile-vector autoregression at different distribution points. GDP increase asymmetrically affects CO2 emissions. This implies that extreme quantiles and the median affect CO2 emissions and economic growth. CO2 emissions and economic growth are bidirectional, says the study. Quantile asymmetry and variance in G7 nations' transmission of effects between variables are also shown. Between 1850 and 2000, CO2 emissions were a net transmitter throughout North America, encompassing the US and Canada. In contrast, (Jebabli et al.,2023) observed that CO2 emissions in Europe (UK, Italy, France, and Germany) and Japan initially acted as net receivers and then as net transmitters.

3. Research Questions (If any)

The research questions corresponding to each of the mentioned study objectives are as follows:

1. How carbon dioxide (CO2) emissions, population size, and economic growth are correlated?
2. Which of the factors, population, or economic growth, exhibits a causal relationship with CO2 emissions?

4. Aim and Objectives

The main objective of this study is to examine the relationship between population growth, economic development, and CO2 emissions from 1850 to 2021. This study aims to investigate the link between economic development and environmental deterioration.

The research objectives are formulated based on the aim of this study which is as follows:

- To analyze and identify the relationship between economic growth and environmental degradation.
- To explore the causality between economic growth and environmental degradation.
- To analyze trends and patterns in the findings of the research.
- To assess the VAR & Multiple Linear regression models and identify the most precise one to find the relationship between economic growth and environmental degradation.
- To use the Vector Autoregressive Model of Timeseries Modeling to forecast CO2 levels based on historical data.

5. Significance of the Study

The primary objective of this study is to examine the correlation between carbon dioxide (CO2) emissions and economic growth through the utilisation of Vector Autoregression (VAR) and Multiple Linear Regression models. To ascertain the primary determinant of carbon dioxide (CO2) emissions in India, it is imperative to investigate the relative influence of two key factors: gross domestic product (GDP) and population size. The study's findings have important policy consequences for the government. Indian Prime Minister Modi has committed to achieving carbon neutrality by the year 2070, while global leaders have collectively agreed to limit the rise in global average temperature to below 2°C. This commitment was made during the COP 26 meeting held in Glasgow. The primary cause of climate change and the rise in global temperatures can be attributed to the significant contribution of greenhouse gas emissions. Carbon dioxide (CO2) emissions make a significant contribution to this phenomenon. To determine the best appropriate machine learning algorithm for our specific use case, we will do a comparative analysis and review relevant literature references.

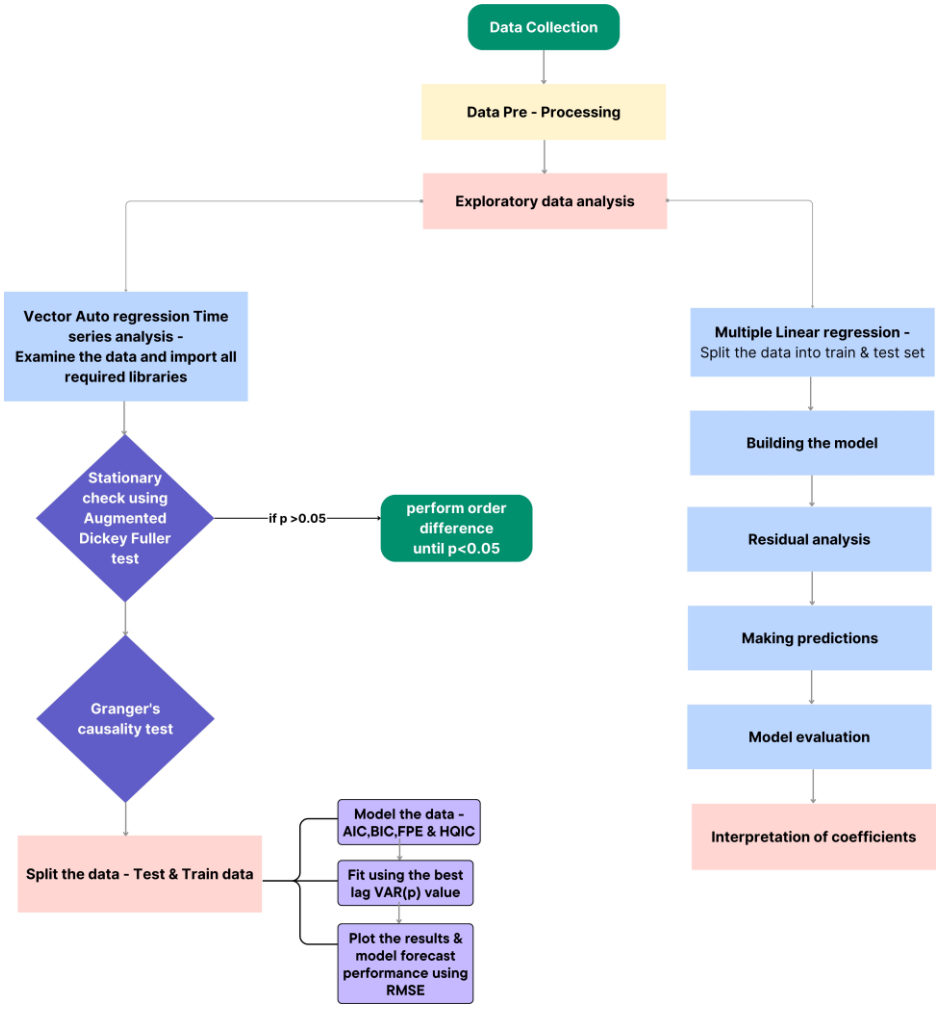
6. Scope of the Study

World in Data dataset on CO2 emissions is used to select the best model to determine the relationship between CO2 emissions and economic growth. The research will make use of data on India's GDP, population, and CO2 emissions including changes in land use from 1850 to 2021. Carbon dioxide (CO2) emissions, which encompass land use change (luc), primarily consider territorial emissions, while neglecting emissions associated with traded goods. Land use change refers to the process by which humans modify the purpose of a given area of land, resulting in its conversion from one use to another. The study's scope is restricted to comparing the VAR model with the Multiple linear regression model, to identify the most accurate model for the study through model evaluation.

7. Research Methodology

One of the objectives is to assess the VAR & Multiple Linear regression models and identify the most precise one to find the relationship between economic growth and environmental degradation. Hence, both the research methodology of VAR & Multiple Linear regression has been explained in detail. The research methodology of the VAR model (list number 7.1 to 7.9) & of Multiple Linear regression (list number 7.10 to 7.15) respectively. The first three steps, namely – Data Collection, Data pre-processing & Exploratory data analysis remain the same in both models. Hence, these three steps have been explained only in the detailed explanation of the VAR model. Figure 3 shows the framework with the workflow of both models.

Figure. 3 VAR & Multiple linear regression Framework illustrating workflow.



7.1 Dataset Description

The dataset, available on the Our World in Data platform, contains 1,705,669 observations with 79 attributes. Among these attributes, three are relevant to the topic: Country, GDP, population, and co2_including_luc. The data covers historical information from 1850 to 2021 for these attributes. Specifically, we are focusing on data related to India for conducting a time series analysis. However, it is important to note that the GDP column has 19% missing data, and the co2_including_luc column has 11% missing data. These missing values can be suitably imputed to facilitate further processing and analysis of the data.

7.2 Exploratory data analysis

19% of GDP missing values and 11% of co2_including_luc missing values will be imputed with mean median or mode values respectively. The changed dataset will be used for further analysis. Read the data and import all the required libraries for the analysis.

7.3 Vector Auto Regression (VAR)

The general form of a vector autoregression (VAR) equation is:

$y\_t = \beta 0 + \beta 1\ y\_\{t-1\} + \beta 2\ y\_\{t-2\} + ... + \beta p\ y\_\{t-p\} + \gamma 1\ x\_\{t-1\} + \gamma 2\ x\_\{t-2\} + ... + \gamma q\ x\_\{t-q\} + \epsilon t$

where:

- yt is a vector of endogenous variables at time t.
- $\beta 0$ is the constant term.
- βi are the coefficients for the lagged values of yt.
- γj are the coefficients for the lagged values of xt, which are exogenous variables.
- ϵt is the error term.

The VAR model can describe the relationship between many quantities over time. The number of endogenous variable lags depends on the VAR model order, p. VAR models with higher orders are more complex and require more data to estimate.

The VAR model has multiple applications, such as variable forecasting and causal connection analysis, identifying system shocks and Understanding system dynamics. The VAR model excels at multivariate time series analysis. It can help you grasp complex interactions between variables and predict future values.

7.4 Check for Stationarity

Prior to using the VAR model, it is imperative to ascertain the stationarity of all variables within the dataset. The concept of stationarity pertains to a statistical characteristic in which a series demonstrates a consistent mean and variance throughout its duration. The Augmented Dickey-Fuller (ADF) test is a frequently used approach for evaluating stationarity. In this experiment, the null hypothesis posits that the time series exhibits non-stationarity. When the p-value obtained from the Augmented Dickey-Fuller (ADF) test is below the significance

level of 0.05, the null hypothesis is rejected. This implies that the time series variables under consideration are deemed to be stationary. To attain stationarity, it is necessary to iteratively use the differencing operation until the data frame exhibits stationarity.

7.5 Granger's Causality Test

Granger's causality test determines variable relationships before modelling. If the p-value of the statistical test is found to be less than 0.05, it can be concluded that there exists a statistically significant relationship between the variables under investigation. Conversely, if the p-value exceeds 0.05, it can be inferred that there is no statistically significant association between the variables. Granger's causality test allows for the identification of both unidirectional and bidirectional causality between the variables.

7.6 Split & Model the Data

Split the dataset into train and test data to model it. Input into the VAR module for modelling purposes. After this, the next step is to select the best-fit lag value. To do this we need to compare different AIC (Akaike information criterion), BIC (Bayesian information criterion), FPE (forecast prediction error) & HQIC (Hannan Quinn Information Criterion). The above parameters will help in the selection of the best-fit lag value.

7.7 Fit the data using the best lag value

The minimum values in combination with AIC, BIC, FPE, and HQIC are given the '*' sign. Select the minimum lag value VAR(p) & fit the model accordingly. Co-efficient, Std. error, t-stat and model probabilities can be seen at every lag till the best-fit lag value.

7.8 Forecast, predict & plot the results

The predicted versus original values of the variables for test data and compare the same to check the accuracy of the forecasted data and actual values metric Root Mean Square error (RMSE) can be used.

7.9 Multiple linear regression

Multiple linear regression is a statistical method used to analyze the correlation between a single dependent variable and multiple independent variables (also known as explanatory variables). The primary aim of multiple regression is to identify a linear equation that can most accurately predict the value of the dependent variable Y based on various values of the independent variables in X.

$y = \beta0 + \beta1x1 + \beta2x2 + \beta3x3 + ... + \beta nxn + \varepsilon$ (Linear Regression: Introduction to Multiple Regression - codingstreets, 2023)

In this equation:

- y stands for the dependent variable (the variable we want to predict).
- β0 represents the intercept (the value of y when all independent variables are zero).
- β1, β2, β3, ..., βn are the coefficients (indicating how much y changes for a one-unit change in each corresponding independent variable x1, x2, x3, ..., xn).
- x1, x2, x3, ..., xn refer to the independent variables.
- ε denotes the error term, which captures the difference between the actual and predicted values of y.

7.10 Spilt the data into the train & test set

The initial fundamental step in regression is to carry out a train-test split. To accomplish this, we employ the sci-kit-learn model. Proper feature scaling plays a crucial role when dealing with numerous independent variables in a model. Since many of these variables may be on different scales, it can lead to obscure coefficients that are difficult to interpret. Scaling the features using the MinMax scaler is essential for two reasons: it enhances the ease of interpretation and facilitates faster convergence of gradient descent methods. With the MinMax scaler, the variables are transformed in a way that all values fall between zero and one, using the maximum and minimum values in the data.

7.11 Building the model

The initial crucial step in constructing the model involves dividing the training data into separate sets, X, and y. Subsequently, the Recursive Feature Elimination (RFE) and linear regression modules from sci-kit-learn are imported. RFE is executed with the desired number of output variables. The process then entails creating the X_test data frame with the RFE-selected variables. Afterwards, a constant variable is added, and the linear model is run using the Ordinary Least Squares method. Features that have a high p-value, indicating that they are not significantly impactful in predicting the dependent variable, are dropped. In regression model building, the null hypothesis corresponding to each p-value posits that the associated independent variable does not influence the dependent variable, while the alternate hypothesis states that it does impact the response. A low p-value (less than 0.05) suggests that the null hypothesis can be rejected. Additionally, redundant features are identified using correlations and Variance Inflation Factor (VIF) >2 and are subsequently dropped. The model is then rebuilt, and the process is repeated iteratively.

7.12 Residual analysis

To perform a Residual Analysis on the train data, we aim to verify whether the error terms exhibit a normal distribution, which is a fundamental assumption in linear regression. To achieve this, we plot a histogram of the error terms and visually examine their distribution. This step is essential in validating the key assumptions of linear regression and ensuring the reliability of the model.

7.13 Making Predictions

After applying scaling on the test sets, we divide them into X_test and y_test. We then utilize our model to make predictions. To do this, we create a new data frame called X_test_new by dropping certain variables from X_test. After adding a constant variable, we proceed to make predictions using the model.

7.14 Model evaluation

The model evaluation involves two steps. First, we visually assess the performance by plotting the actual y_test values against the predicted y_predicted values. This comparison allows us to observe how well the model's predictions align with the actual data. Second, we utilize the R^2 value to evaluate the model's performance on the test data. The R^2 value provides a measure of how well the model fits the test data and indicates the proportion of the variance in the dependent variable that can be explained by the independent variables.

7.15 Interpretation of coefficients

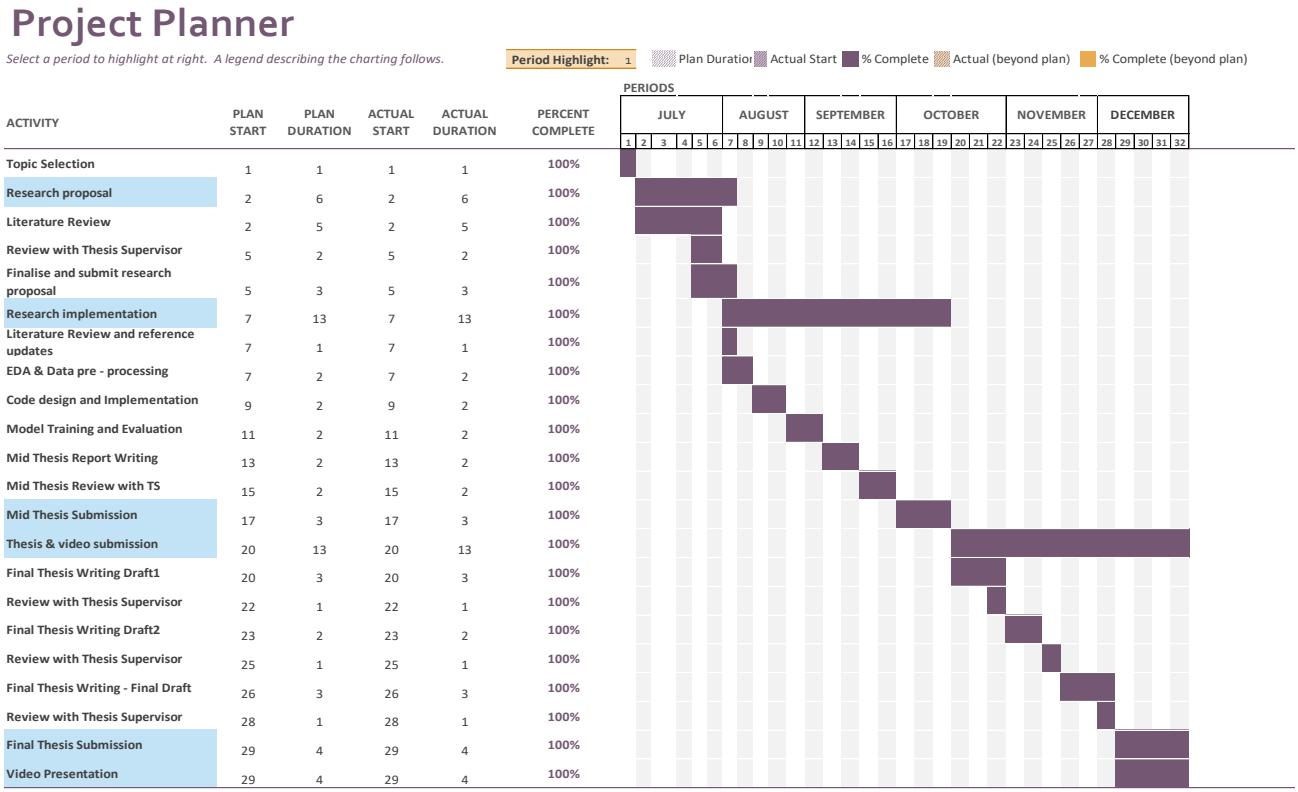
Interpretation of coefficients obtained from the last iteration of OLS regression results and identify the factors which contribute positively towards the explanation of the increase in CO2 levels.

8. Requirements Resources

- Hardware
  - Min i7 9th Gen Windows PC with 16GB RAM
  - Jupyter Notebook 6.4.5
- Software
  - Operating System: Windows 11
  - Python version 3.9.7, MS Office 365
  - Python libraries - NumPy, Matplotlib, Seaborn, Pandas, pylab, Scikit learn.
- Others
  - Access to journals, research papers, and conference papers (IEEE, Elsevier, Springer etc.

9. Research Plan

Figure 4. Gantt Chart – Project plan and execution



ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Topic Selection	1	1	1	1	100%
Research proposal	2	6	2	6	100%
Literature Review	2	5	2	5	100%
Review with Thesis Supervisor	5	2	5	2	100%
Finalise and submit research proposal	5	3	5	3	100%
Research implementation	7	13	7	13	100%
Literature Review and reference updates	7	1	7	1	100%
EDA & Data pre - processing	7	2	7	2	100%
Code design and Implementation	9	2	9	2	100%
Model Training and Evaluation	11	2	11	2	100%
Mid Thesis Report Writing	13	2	13	2	100%
Mid Thesis Review with TS	15	2	15	2	100%
Mid Thesis Submission	17	3	17	3	100%
Thesis & video submission	20	13	20	13	100%
Final Thesis Writing Draft1	20	3	20	3	100%
Review with Thesis Supervisor	22	1	22	1	100%
Final Thesis Writing Draft2	23	2	23	2	100%
Review with Thesis Supervisor	25	1	25	1	100%
Final Thesis Writing - Final Draft	26	3	26	3	100%
Review with Thesis Supervisor	28	1	28	1	100%
Final Thesis Submission	29	4	29	4	100%
Video Presentation	29	4	29	4	100%

Figure. 4 is the Gantt Chart which represents the planning and execution part of the Thesis duration.

Gantt chart_
Timeseries_Chetan_Tip

References

Abbasi, K.R., Kirikkaleli, D. and Altuntaş, M., (2022) Carbon dioxide intensity of GDP and environmental degradation in an emerging country. *Environmental Science and Pollution Research*, 2956, pp.84451–84459.

Adebayo, T.S., Awosusi, A.A., Kirikkaleli, D., Gbenga, &, Akinsola, D., Madhy, &, Mwamba, N., Tomiwa, *, Adebayo, S., Akinsola, G.D. and Mwamba, M.N., (n.d.) Can CO 2 emissions and energy consumption determine the economic performance of South Korea? A time series analysis. [online] Available at: https://doi.org/10.1007/s11356-021-13498-1.

Ali, S.S.S., Razman, M.R. and Awang, A., (2020) The nexus of population, gross domestic product growth, electricity generation, electricity consumption and carbon emissions output in Malaysia. *International Journal of Energy Economics and Policy*, 103, pp.84–89.

Chang, T., Hsu, C.M., Chen, S.T., Wang, M.C. and Wu, C.F., (2023) Revisiting economic growth and CO2 emissions nexus in Taiwan using a mixed-frequency VAR model. *Economic Analysis and Policy*, 79, pp.319–342.

Hao, Y., (2022) The relationship between renewable energy consumption, carbon emissions, output, and export in industrial and agricultural sectors: evidence from China. *Environmental Science and Pollution Research*, 2942, pp.63081–63098.

Jebabli, I., Lahiani, A. and Mefteh-Wali, S., (2023) Quantile connectedness between CO2 emissions and economic growth in G7 countries. *Resources Policy*, 81.

Khan, M.K., Khan, M.I. and Rehan, M., (2020) The relationship between energy consumption, economic growth and carbon dioxide emissions in Pakistan. *Financial Innovation*, 61.

Nguyen, V.C.T. and Le, H.Q., (2022) Renewable energy consumption, nonrenewable energy consumption, CO2 emissions and economic growth in Vietnam. *Management of Environmental Quality: An International Journal*, 332, pp.419–434.

Osobajo, O.A., Otitoju, A., Otitoju, M.A. and Oke, A., (2020) The impact of energy consumption and economic growth on carbon dioxide emissions. *Sustainability (Switzerland)*, 1219.

Pandey, K.K. and Rastogi, H., (2019) Effect of energy consumption & economic growth on environmental degradation in India: A time series modelling. In: *Energy Procedia*. Elsevier Ltd, pp.4232–4237.

Puntoon, W., Tarkhamtham, P. and Tansuchat, R., (2022) The impacts of economic growth, industrial production, and energy consumption on CO2 emissions: A case study of leading CO2 emitting countries. *Energy Reports*, 8, pp.414–419.

Salari, M., Javid, R.J. and Noghanibehambari, H., (2021) The nexus between CO2 emissions, energy consumption, and economic growth in the U.S. *Economic Analysis and Policy*, 69, pp.182–194.

Singh, K. and Kaur, J., (2022) Do energy consumption and carbon emissions impact economic growth? New insights from India using ARDL approach. *OPEC Energy Review*, 461, pp.68–105.

Udemba, E.N., Güngör, H., Bekun, F.V. and Kirikkaleli, D., (2021) Economic performance of India amidst high CO2 emissions. *Sustainable Production and Consumption*, 27, pp.52–60.

Wang, C.M., Hsueh, H.P., Li, F. and Wu, C.F., (2019) Bootstrap ARDL on Health Expenditure, CO2 Emissions, and GDP Growth Relationship for 18 OECD Countries. *Frontiers in Public Health*, 7.

Washington, DC: World Resources Institute, (2022) | *Greenhouse Gas (GHG) Emissions* | *Climate Watch*. [online] | Greenhouse Gas (GHG) Emissions | Climate Watch. Available at: https://www.climatewatchdata.org/ghg-emissions?chartType=percentage&end_year=2020&gases=co2&start_year=1990 [Accessed 13 Aug. 2023].

Anon (2023) Linear Regression: Introduction to Multiple Regression - codingstreets. [online] Linear Regression: Introduction to Multiple Regression - codingstreets. Available at: https://codingstreets.com/linear-regression-introduction-to-multiple-regression/ [Accessed 19 Aug. 2023].