UCLouvain

LOUVAIN
School of Management

Louvain School of Management

Forecasting fossil CO2 emissions in China and India: an analysis with various predictors

Author(s): Sébastien Wéry
Supervisor(s): Pr. Leonardo Iania
Academic year 2023-2024
Dissertation for the master of Business Engineering
Master subject and focus: Energy management
Daytime schedule/Staggered schedule

During the preparation of this master's thesis, the author(s) utilized the following AI tools:

1. Grammarly

Used for checking grammar mistakes and spelling. All parts of the thesis are affected.

2. ChatGPT

Used for language and structure improvement, as well as coding content. All parts of the thesis are affected.

3. Scribbr

Used to generate citations and references with titles as input. The part "Bibliography" is affected.

After using the tools mentioned above, the author(s) diligently reviewed and edited the content produced by the tools. I take full responsibility for the final content presented in this thesis.

By signing this declaration, I affirm that the content of this master's thesis reflects my original work, augmented by the responsible use of AI.

The 3rd of August 2024,

Sébastien Wéry,

At first, I would like to thank my master's thesis supervisor, Pr. Leonardo Iania, for the time, energy, and precious advice that helped me realize this thesis over the last 12 months. The subject that Prof. Iania suggested was very interesting and rewarding to work on. I would also like to thank the Université Catholique de Louvain and the teachers I've had over the past five years. The completion of this thesis wouldn't have been possible without the high-quality education I received during this time. Additionally, I would like to give special thanks to Prof. Betrand Candelon for the class "Forecasting", which was very useful for this paper. Last but not least, I would like to thank my family, and more especially my mother and my father for the insightful discussions we've had on self-attitude, motivation, and organisation in writting a master's thesis. The conversations regarding the purposes of this paper and the direction I should take when collecting and summarizing information were truly helpful.

Abstract

In this paper, the fossil emissions of China and India are forecasted in the medium term (2025) using ARIMA model and long-term (2030) using VECM model. It is shown that the fossil emissions of both countries are expected to increase by a factor of 1,21 for China and 1,80 for India between 2022 and 2030. This research also studies the causal relationship between 33 macroeconomic indicators and China's fossil CO2 emissions between 1961 and 2022, and between 35 macroeconomic indicators and India's fossil CO2 emissions between 1961 and 2022. Consistency is shown with the analyses of Bennedsen et al. (2021), where predictors related to economic activity and production have, on average, the highest R-squared values among all predictors. Finally, using CO2 VECM forecasts and S&P GDP forecasts, it is shown that both China and India are not expected to meet their carbon intensity targets for 2030 if they do not change their methods of energy and goods production.

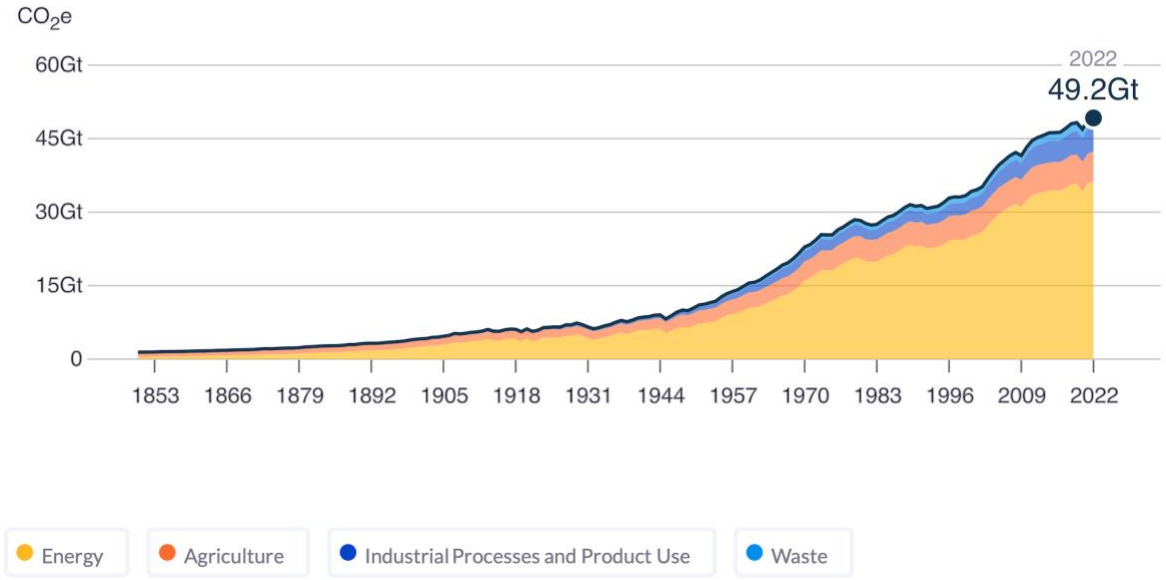
UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve
Boulevard Emile Devreux 6, 6000 Charleroi, Belgique
Chaussée de Binche 151, 7000 Mons, Belgique
www.uclouvain.be/lsm

Table of Contents

1. Introduction

Greenhouse gases are necessary to support life on Earth. Without the presence of greenhouse gases in the atmosphere, the average surface temperature on Earth would be around -18°C (Yoro & Daramola, 2020). These gases are caused by both natural procedures and human activity. While this is true, their concentration in earth atmosphere has strongly increased since the second industrial revolution (Yoro & Daramola, 2020), mainly due to the production of electricity and energy. This is observable on the chart below (Climate Watch, 2024):

Worldwide GHG emissions since 1850



After fuels are burned, and through a chemical reaction, they reject GHG in earth atmosphere. When greenhouse gases absorb solar radiations, they prevent these radiations to go back to space and maintain high temperatures in the low atmosphere (United Sates Environmental Protection Agency, 2024). These increases in temperatures result in major worldwide issues that fall under the broader category of "climate change" (Yoro & Daramola, 2020).

In 2022, China and India emitted 14.4 and 3.52 GtCO2eq of GHG respectively, with CO2 comprising 78.4% of India's total and 82.6% of China's total (Gütschow & Pflüger, 2023). These two countries are also the 1st and 3rd largest worldwide GHG emitters since 2008 (European Commission, 2022), with China representing 29.1% and India 7.33% of the world GHG emissions. Additionally, the CO2 emissions of India should only peak between 2040 and 2045 and the GHG emissions would increase by 41% between 2023 and 2030 (Chateau, 2023). Regarding China, the yearly CO2 emissions should continue to increase until 2030, before ideally starting to decrease from that year (International Energy Agency, 2023).

Overall, these two countries are critical actors in the environmental transition, and forecasting their CO2 emissions would allow governments and environmental organizations to understand the direction both countries are taking, and to adapt to it. The number of relevant articles and available data for forecasting CO2 emissions in India and China also justified this research, as there is much less information compared to the U.S. and most European countries. This gap highlights the relevancy of analyzing how India's and China's CO2 emissions are projected to evolve through time, but also to determine whether the publicly available data is sufficient for making reliable forecasts.

Therefore, the fossil CO2 emissions of India and China are forecasted using different methods. First, emissions are forecasted between 2023 and 2025 using the auto-regressive integrated moving-average (ARIMA) model. The purpose of this first model is to compare its forecasting accuracy with a more advanced model, the vector error correction model (VECM), to determine whether including external variables enhances accuracy. The VECM is used to forecast fossil CO2 emissions of both countries until 2030 and to compare these results with the emissions targets of both countries.

In addition to the difficulty of finding relevant information and reliable data to forecast the emissions of India and China, I faced other challenges. I highlight the time-consuming aspect of data preparation and the learning of complex forecasting algorithms.

2. Literature Review

2.1. The origin of GHG and CO2 emissions, and their impact on the environment

There are seven types of greenhouse gases: water vapor (H2O), carbon dioxide (CO2), nitrous oxide (N2O), methane (CH4), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and sulfur hexafluoride (SF6) (Jones et al., 2023). All cause the greenhouse effect, but not for a similar period and at a similar scale. Indeed, CO2 gases can stay several decades or centuries in the atmosphere. (European Commission, 2024). This is why the term "CO2eq", or "equivalent greenhouse gas emissions in carbon dioxide" is utilized in the scientific literature, and will be used in this research document, to gauge the emissions of a specific economy.

CO2, CH4, and N2O are gases the most responsible for greenhouse effect. In 2021, CO2, CH4, and N2O represented respectively 75%, 5%, and 19% of worldwide emitted GHG (Jones et al., 2023). Although water vapors are naturally the most present GHG in the atmosphere, CO2 is largely the most diffused in the atmosphere (Yoro & Daramola, 2020).

CO2 emissions are caused by the combustion reaction of fossil fuels such as coal and oils, but also solid waste, such as trees and other biological materials. It is also the result of certain chemical reactions, such as cement manufacturing and plastic production (United States Environmental Protaction Agency, 2023). Overall, on a sectorial analysis, CO2 emissions are mainly due to the production of electricity, food, and goods through the combustion of fossil fuels, but also caused by the transportation of people and goods (United Nations, 2024). Additionally, the absorption capacity of CO2 emissions is impacted by deforestation (European Union, 2024). As trees and forests play an important role in CO2 absorption, removing these biological organisms has a direct impact on CO2 levels.

CH4 emissions are the second most important human-influenced greenhouse gas, after carbon dioxide. Out of the human related emissions, “agriculture and waste emissions”, “fossil fuel emissions”, and “biomass and biofuel burning emissions” are highlighted (Saunois et al., 2024).

N2O emissions are mainly due to agricultural activities, such as the use of nitric oxide fertilizers, to the combustion of fossil fuels, and to wastewater management (United States Environmental Protaction Agency, 2023).

2.2. The reasons behind forecasting CO2 emissions

In general, many reasons can justify the forecasting of CO2 emissions. The most relevant ones give credibility to this research.

First of all, it is useful for a government to create a coherent greenhouse gas emissions policy plan, i.e. a plan that contributes to the limitation of global warming in the long run, while softening accordingly the social cost of carbon dioxide. The social cost of carbon dioxide is defined as the dollar value of climate change damages following a marginal increase in CO2 emissions (Rode et al., 2021). By forecasting accordingly variations in CO2 emissions for a country, governments can take better-informed decisions regarding their greenhouse gas policy. To illustrate this, if a country would see its CO2 emissions sharply increase but didn't forecast it accordingly, the additional costs of having no time buffer for implementing new environmentally related projects could be significant. Additionally, countries that have specific objectives in reducing their CO2 emissions must make some challenging social decisions. For instance, the EU relies on a vast portfolio of policy instruments to become climate-neutral by 2050. Such economic, educational, or financial measures have direct consequences on the standard of living (European Environment Agency, 2021). Overall, emissions forecasts allow governments to make better-informed decisions, financially and socially.

Secondly, it is necessary to foresee a country's expected level of CO2 emissions to compare it with the CO2 targets decided in the Paris Agreement (Le Quéré et al., 2018). This further allows governments to know whether their expected GHG or CO2 emissions are aligned with Paris Agreement. To obtain expected levels of GHG or CO2 emissions that are reliable, statistical forecasting methods are necessary. For instance, the Global Carbon Budget develops different statistical models, which allows it to compare both estimated CO2 emissions previsions and targeted CO2 emissions for different countries (Friedlingstein et al., 2022).

Thirdly, evaluating correctly CO2 concentrations and their repartition in the atmosphere can help to better understand the global carbon cycle (Friedlingstein et al., 2022). As this research paper tries to evaluate future CO2 emissions for two countries using macroeconomic, social, and environmental variables, it also studies the causal relationships between diverse variables and CO2 emissions. Such outcomes could help to understand better the causes of anthropogenic CO2 emissions and the effect of it on the global carbon cycle.

2.3. Literature review on forecasting CO2 emissions

In this chapter, I look deeper into the matter of CO2 emissions forecasting by reviewing previous relevant studies. First, I look at the different types of statistical methodologies already used to forecast CO2 and GHG emissions. To forecast time series in statistics, one can use either only the past information about a variable, i.e., its past values, or one can include some "external" variables if a statistical correlation in the trend evolution is observed between the variable of interest and another variable.

2.3.1. Univariate literature for forecasting CO2 emissions

A first method very often used to forecast univariate time series, which incorporate no external variable, is the auto-regressive integrated moving-average (ARIMA) model (Hyndman and Athanasopoulos, 2013). ARIMA (p,d,q), is a model where “p” is the number of auto-regressive processes, “d” is the order required to reach stationarity, and “q” is the number of moving average processes (Box and Jenkins, 1970). While the auto-regressive part uses only the last values of a time series, the moving-average part uses the stochastic error made by the model for past forecasts (Box and Jenkins, 1970). Rahman and Hasan (2017) forecasted yearly CO2 emissions in Bangladesh between 2016 and 2018, using the CO2 emissions for the period 1972-2015. They found out that the best ARIMA model for this time series was an ARIMA (0,2,1). The method used by the authors, which consists in computing and comparing various error-minimization criteria to choose the best model, proved to be efficient (Kumari et al., 2014). Ning et al. (2021) also used an ARIMA model to forecast the CO2 emissions of four regions in China (Beijing, Henan, Guangdong, and Zhejiang) between 2018 and 2020, using the yearly CO2 emissions between 1997 and 2017. The authors used the Akaike information criteria (AIC) (Akaike, 1973) and Schwartz criteria (SC) (Schwarz, 1978) which compute the forecasting errors in two different ways, as well as the R-squared criteria, which indicates to what extent the variability of the original time series is explained. The optimal models found for the four regressions are ARIMA (2,2,0) for Henan and Beijing regions, ARIMA (1,0,0) for Guangdong, and ARIMA (3,2,4) for the Zhejiang region. Regarding the level of CO2 emissions forecasted for the period 2018-2020, there is no clear consensus. It was expected to decrease of 3% per year for two of the four Chinese cities, while it was the opposite for the last two cities. The R-squared values found for the four time-series using the best-fitting ARIMA models vary between 0,3 and 0,94, which the authors qualify as between "good" and "very good" forecasting results.

Some authors used others univariate methods to forecast CO2 emissions. Tudor (2016) used seven univariate forecasting models to predict Bahrain's yearly per capita CO2 emissions between 2012 and 2021. Using the CO2 emissions of the country between 1960 and 2011 and the root mean squared error (RMSE) criteria, the author found that the best-performing model was the neural network auto-regressive model (NNAR)[1]. On average, the ARIMA model had forecasting results that are 47 percent worse than the NNAR model, but it still performed better than the structural time series (STS) models and two others based on the exponential smoothing method[2] (Holt-winters and TBATS). Additionally, Fatima et al. (2019) analyzed the CO2 emissions of 9 Asian countries using ARIMA and simple exponential smoothing (SES) models. Using yearly CO2 emissions between 1971 and 2014, they showed that the SES model, which is the weighted linear sum of recent past data points (Muth, 1960), has a higher forecast mean absolute errors (FMAE) than the ARIMA model for 6 of the 9 countries. In this case, the ARIMA model fitted the data better than the SES model. Köne and Büke (2010) used more classical univariate forecasting methods. The authors used the linear regression method and forecasted the CO2 emissions of 25 countries for 2015 and 2030. Using this method, which forecasts time series linearly by using the past values only (Köne & Büke, 2010), the authors got significant statistical trends for 11 countries out of the 25. India is one of the 11 countries, and they forecasted that Indian CO2 emissions should increase by 32.7% between 2015 and 2030.

Overall, even if it doesn't seem that ARIMA models have the best forecasting accuracy among univariate models, they are easy to implement and are widely used in the literature for quality comparison purposes with more advanced models.

2.3.2. Multivariate literature for forecasting CO2 emissions

I am now focusing on multivariate time series analysis. I examine articles incorporating external variables, such as economic growth and energy usage, to predict or model CO2 emissions. A first method is to use auto-regressive distributed lag (ARDL) to test the stationarity of a time series, and then to use the vector error correction model (VECM) technique to test short and long-term causality between variables (Boontome et al., 2019). Aftab et al. (2021) used this method to estimate relationships between CO2 emissions and

[1] NNAR model is an artificial network forecasting time series using only one hidden layer and inputs that can be lagged (Tudor, 2016).

[2] The exponential smoothing method, which is widely used for forecasting purposes (Hyndman and Athanasopoulos, 2013), applies weighted averages of past observations with exponentially decreasing weights to highlight longer-term trends and cycles (Hyndman and Khandakar, 2008).

diverse macroeconomic and environmental variables in Bangladesh between 1971 and 2019. They showed that energy consumption, gross domestic product (GDP), urbanization, forest area, and research and development have a positive long-term causal relationship towards CO2 emissions. In the short term, urbanization and forest area have a negative causal relationship with CO2 emissions. Boontome et al. (2019) also used the ARDL and VECM methodology to find relationships between CO2 emissions and diverse economic and energy variables in Thailand between 1990 and 2016. As well as Aftab et al. (2021), they found that GDP and energy consumption have a positive correlation in the long term with CO2 emissions, whereas oil prices and renewable energy consumption have an opposite effect. They also found out that in the short term, a decrease in oil prices lead to higher CO2 emissions and that an increase in renewable energy usage led to a decrease in both energy consumption and CO2 emissions, which make sense given the increase in renewable energy usage over the past years for Thailand (Boontome et al., 2019). Aboul Ela (2023) forecasted Egypt's CO2 and CH4 emissions between 2020 and 2030 using VECM model. The author also found out that energy consumption and GDP have strong impact on CO2 emissions, and that international trade and Agricultural production have the same positive effect. In contrast, there was no clear sign of a correlation between CO2 emissions and foreign direct investment, and commodity trade.

I also give interest in other multivariate forecasting models. Aboul Ela (2023) used the autoregressive integrated moving average with external variables (ARIMAX) model to compare its performance with VECM model in forecasting Egyptian CO2 emissions. Using diverse error related criteria, the author observed that the VECM model had much more precision in forecasting CO2 emissions as it has lower values than ARIMAX model on each criterion.

Bennedsen et al. (2021) modeled and forecasted the annual CO2 emissions of USA between 1960 and 2017 using the structural augmented dynamic factor model (SADFM) and 226 different macroeconomic, agricultural, and foreign trade variables. The authors noticed that SADFM model predicted better USA CO2 emissions than the ARIMA model. Using the Lasso technique, which retains the best sets of variables relevant together (Tibshirani, 1996), they found that the variables helping to model CO2 emissions the best are real economic variables such as industrial production and household consumption, employment variables, and trade industry sales. These results may suggest that very specific economic variables could be more relevant than the GDP in capturing forecasting information. Algieri et al.

(2023) also found that using more specific environment-related variables, rather than total energy consumption, could improve modeling and forecasting results. The authors modeled the quarterly CO2 emissions of USA between 1971 and 2021 and forecasted it between 2021 and 2023 using thirteen forecasting models mostly related to quantiles linear combination. To do so, they tested the correlation between quarterly USA CO2 emissions and the 248 variables of the McCracken and Ng (2020) database involving financial, environmental, economic, and social indicators, as well as additional variables related to economic activity collected through surveys by Michigan University, 15 variables linked to energy consumption, production, prices, import, and export, and finally, some variables related to nature such as average US temperature anomalies and drought events. The authors concluded that selecting a broader range of variables than GDP and total energy consumption is very beneficial for modeling and forecasting CO2 emissions. Indeed, they show in their study that implementing and combining macroeconomic drivers, nature-related indicators such as drought and temperature anomalies variables, and energy factors improves forecasting quality. They also found that mostly all the models built with these very specific variables related to economy, energy, environment, and nature outperformed the ARIMA (4,0,0) model with an additional R-squared value of 30%.

Another type of method often used in the literature is neural network algorithms. Overall, neural networks are self-trainable algorithms able to make forecasts based on available data (Kohzadi et al., 1995). Faruque et al. (2022) forecasted CO2 emissions of Bangladesh for 2020 and 2021 using the yearly CO2 emissions, GDP, and energy consumption of the same country between 1972 and 2019. Using the fully modified ordinary least squares (FMOLS) method, they proved that in the long run, there is a strong positive correlation between energy consumption and CO2 emissions and a small negative correlation between GDP and CO2 emissions. In their study, they used four deep learning algorithms to forecast CO2 emissions in Bangladesh: the convolution neural network (CNN), the CNN long short-term memory (CNN–LSTM), the long short-term memory (LSTM), and the dense neural network (DNN). Even if the models developed by the authors have low error-related criteria (MAPE below 5%), there are limitations to forecasting with neural network algorithms. Indeed, Kohzadi et al. (1995), highlighted some issues with neural network models. First, due to their non-linear nature, neural networks need large datasets and consequently many parameters to produce accurate forecasts. Secondly, and most importantly, neural networks cannot identify the contribution of each variable to forecasting results, which results in a lack of interpretability of the model (Kohzadi et al., 1995). Jena et al. (2021) also used a derivative of neural

networks, the artificial neural network (ANN), to forecast the CO2 emissions of 17 large countries at the global level. According to the authors, CO2 emissions were expected to increase by 0.05% per year in China and 0.58% per year in India between 2017 and 2019. They also state that the prediction error of their neural network model was high for forecasting CO2 emissions of several countries.

Overall, based on the reviewed literature, using the VECM model for forecasting purposes offers a well-balanced approach between accuracy and interpretability.

3. Data and Methodology

This part of the research will discuss the methodology and data used to forecast India and China's fossil CO2 emissions in the short and mid-term.

Overall, for each model, the methodology consists of the following steps: first, the data is collected and transformed by cleaning it, making it stationarity, and removing the outliers. It ensures the data is usable for forecasting purposes. Then, for the VECM model only, the most relevant predictors are selected for the forecast. After this, the ARIMA or VECM model is built. I then check if the residuals resulting from it are white noise. Finally, I forecast the time series using the model previously built and evaluate the forecasting quality using several metrics.

3.1. Data collection

3.1.1. CO2 emissions

The yearly fossil CO2 emissions of China and India from 1961 to 2022 were both retrieved from the Global Carbon Atlas website[3], which posts yearly updates of CO2 emissions according to the Global Carbon Project (GCP)[4] methodology.

Andrew and Peters (2022) explain that the GCP data[5] does not directly measure CO2 emissions level in the atmosphere directly; instead, these emissions are estimated based on the amount and of fossil fuels (solid, liquid, and gaseous) consumed, gas flaring, and cement production by sector of activity. Because the GCP prioritizes accuracy over the use of a uniform method across all countries, different methodologies are applied for each country. However, the general methodology involves combining data from CDIAC-FF, BP, and national sources to build their datasets.

For both India and China, the baseline of the dataset is built by the CDIAC-FF (Carbon Dioxide Information Analysis Center - Fossil Fuels). This dataset includes two sources itself

[3] https://globalcarbonatlas.org/emissions/carbon-emissions/

[4] The Global Carbon Project was established in 2001 by a shared partnership between the International Geosphere-Biosphere Programme (IGBP), the International Human Dimensions Programme (IHDP), the World Climate Research Programme (WCRP) and Diversitas. The scientific goal of the Global Carbon Project is to develop a complete picture of the global carbon cycle, including both its biophysical and human dimensions together with the interactions and feedbacks between them (GCP, 2024).

[5] The dataset from the Postdam Institute for Climate Change (PIK) could also have been used for this research paper. However, due to the methodologies employed by PIK, the comparability of emissions across different times and countries is limited (Climate Watch, 2024), making the GCP dataset to seem more reliable.

(Andrew and Peters, 2022). Data concerning production, processing, trade, and consumption of fossil fuels are assembled annually by the United Nations (UN) Statistics Division.[6] Data from the United States Geological Survey are the foundation of the CDIAC-FF estimates of emissions for limestone calcination and cement manufacture. BP Statistical Review is also used in the GCP to extend the CDIAC-FF data in recent years to China and India, where UN data are lagged. Finally, additional data from UNFCCC (United Nations Framework Convention on Climate Change) and other national reporting bodies are used to refine and adjust the baseline data (Andrew and Peters, 2022). For China's CO2 emissions, the annual estimates incorporate detailed data from national sources for cement and lime production. For India's CO2 emissions, the national fiscal year reporting system (historical data from the Indian Bureau of Mines and Ministry of Coal) is used to re-estimate fossil CO2 emissions, by offering a breakdown of CO2 emissions by sources (coal, oil, cement, natural gas). After estimating the annual quantities of fossil fuels consumed, gas flaring, and cement manufacture, conversion factors are applied by the GCP to determine the resulting estimated yearly fossil fuel emissions. The computations can be represented with the following formula:

$$CO2_{\,i,t} = \sum_{x=1}^{k} fuel_{x,i,t} \,.\, k_x$$

where *x* represents the type of fuel or product emitting CO2, *i* is the index for the country, and *t* is the year studied. "*CO2*" denotes the estimated fossil fuel emissions in megatons, "*fuel*" indicates the amount of each fuel or gas flared consumed, or cement produced, and "*k*" is a constant conversion factor for each fuel type, gas flaring, or cement production, expressed in megatons of CO2 per megaton of fuel, gas flared, or cement.

3.1.2 Explanatory variables

These variables will be used for the multivariate forecasting section. The whole dataset was collected on the world bank data website, under the section "world development indicators"[7]. For each country, there are theoretically 1492 available indicators, but only few of them are studied for China or India, and very rarely for the period covered in this research paper, which is between 1961 and 2022. Out of these possible indicators to use, 33 are selected to test the

[6] The UN Statistics Division utilizes apparent consumption of fossil fuels, calculated from national accounts covering production, exports, imports, and stock variations. This differs from observed consumption, gathered from non-traditional sources such as industry reporting (Andrew and Peters, 2022).
[7] https://databank.worldbank.org/source/world-development-indicators

fit with China's fossil CO2 emissions and 35 with India's fossil CO2 emissions. These variables are available in Appendix 1.

Both countries share the same 33 variables, but two additional variables are used for India and not for China. Bennedsen et al. (2021) showed very good correlation (high R-squared value) between "exports as a capacity to import" and CO2 emissions, and "inflation in consumer prices" and CO2 emissions. I decided to include these two additional variables for India, but they aren't available for China. Additionally, it is important to note that the variables related to fuel consumption aren't selected here to avoid obvious multicollinearity risks, as the CO2 emissions dataset is mainly based on fuel consumption data.

These world development indicators are compiled by the World Bank. The organism directly or indirectly[8] collaborates with the national statistical departments of each member state, as well as national important organizations, such as banks, to get the datasets baseline. Consequently, the quality of national data depends on the reporting quality of these entities. Nevertheless, the General Data Dissemination System (GDDS) and the Data Quality Assessment Framework (DQAF), both developed by the IMF in collaboration with the World Bank, provide frameworks for assessing national statistical systems and data quality. Overall, the World Bank's methodology for developing World Development Indicators involves gathering various information sources, using GDDS and DQAF to select relevant and trustworthy ones, and combining them if necessary (The World Bank, 2024). Precise data sources are available in the excel annexed to this paper, or at the World bank website mentioned above.

3.2. Methodology for ARIMA modeling

3.2.1. Data cleaning and importation

After the data is collected, the yearly fossil CO2 emissions of China and India from 1961 to 2022 are imported in RStudio for analysis. The dataset containing world development indicators doesn't need to be imported yet, as ARIMA modeling is univariate.

[8] Via partners (The world bank, 2024): the United Nations, the Organisation for Economic Co-operation and Development (OECD), and the International Monetary Fund (IMF).

3.2.2. Time series stationarity

After the data is imported, I create a time series for both China's and India's fossil CO2 emissions, indicating that the variables are associated to a specific time period (1961 to 2022).

Following this, the first step to do in a regression is to check the stationarity of the variables. Stationarity is crucial in time series forecasting because many statistical models assume that the properties of the series do not change over time. Non-stationary data lead to unreliable and spurious results. Stationarity implies that the mean, variance, and autocorrelation structure of the series remain constant over time (Dickey and Fuller, 1979). To test for stationarity, augmented Dickey-Fuller (ADF) test is firstly used. The ADF test (Dickey and Fuller, 1979) is used to test for a unit root in a time series. The presence of a unit root indicates that the time series is non-stationary. The ADF test first involves estimating the following regression:

$$\Delta Y_t = \alpha + \beta t + \gamma Y_{t-1} + \sum_{i=1}^{p} \delta_i \Delta Y_{t-i} + \epsilon_t$$

where $\Delta Y_t = Y_t - Y_{t-1}$, is the first difference of the series, α is the drift term, β is the time trend, γ is the coefficient of the lagged level of the series, δ_i are the coefficients of the lagged differences of the series, p is the number of lagged difference terms, ϵ_t is the error term. The test statistic for the ADF test is the t-statistic for the coefficient γ, where γ' is the estimated value of γ and SE(γ') is the standard error of the estimate:

$$\tau = \frac{\gamma'}{SE(\gamma')}$$

The two hypotheses of the test are the following:

- H_0 (null hypothesis): The time series is non-stationary (i.e., has a unit root it);
- H_1 (alternative hypothesis): The time series is stationary (i.e., it does not have a unit root).

After comparing the test statistic τ to the critical values from the augmented Dickey-Fuller distribution, if τ is less than the critical value of 5% (arbitrary decision), I reject the null hypothesis H0 and conclude that the series is stationary. If the null hypothesis isn't rejected, the time series is not yet stationary.

To reach stationarity, some techniques are possible to use. Two are used in this research paper. The first one is to take the logarithms of the time series. It reduces the difference between values of a variable, which increases the probability to transform a non-stationary time series into a stationary one. Another method is to differentiate the time series. Differencing a time series is to look at the difference between two values of a variable separated by one period, which is one year in this cas. A time series can be differentiated several times, but it removes a degree of freedom at each differentiation, because the size of the time series (n) is reduced by one at each step. If the time series isn't yet stationary (according to the ADF test) after taking the logarithms and after having taken a first order of differentiation, the Philipps-Peron test (Phillips and Peron, 1988) is used. This test also looks for a unit root in a time series, but it makes different assumptions about the error term. It is more robust to general forms of heteroscedasticity in the error term. The PP test is based on the following autoregressive model:

$$Y_t = \alpha + \rho Y_{t-1} + \epsilon_t$$

And has the following hypotheses:

- $H_0 : \hat{\rho} = 1$, the time series is non-stationary (i.e., has a unit root it);
- $H_1 : \hat{\rho} < 1$, the time series is stationary (i.e., it does not have a unit root);

where $\hat{\rho}$ is the estimated value of ρ, which is the autoregressive parameter in the time series. After computing the usual t-test statistic, if it is less than the critical value of 5%, I reject the null hypothesis H_0.

After a first order of differentiation, if both ADF and PP tests goes into the direction that the time series is not stationary, I differentiate it a second time. If the ADF test indicates that the time series is not stationary, but the PP test suggests otherwise, it is possibly due to the presence of structural breaks (Bai and Perron, 1998). These occur when there is an unexpected and permanent change in the parameters of regression models, which can impact the results of the ADF test. Then, I use the Chow-test (Chow, 1960) to check if there is a structural break at a known point in a time series. This could be due to policy changes, economic events, or other external factors. The test has the following hypothesis:

- H_0: There is no structural break at the specified point.
- H_1: There is a structural break at the specified point.

The Chow Test involves splitting the dataset at the suspected break point Tb and estimating the following regression models. The break point (at $t = T_b$) is chosen after looking at the evolution of the time series across time and could be present when the value of the variables suddenly changes. For the first and second segment, respectively:

$$Y_t = \alpha_1 + \beta_1 X_t + \epsilon_{1,t} \quad \text{for } t = 1,2,\dots,T_b \; ;$$
$$Y_t = \alpha_2 + \beta_2 X_t + \epsilon_{2,t} \quad \text{for } t = T_b + 1,\dots,T_b + 2,\dots,T$$

The Chow Test statistic FF is calculated as:

$$F = \frac{\big(RSS_1 - (RSS_2 + RSS_3)\big)/k}{(RSS_2 + RSS_3)/(n_1 + n_2 - 2k)} \sim F_{k,(n_1+n_2-2k)}$$

where: RSS_1, RSS_2, RSS_3 are the residual sum of squares for the combined dataset, for the first segment, and for the second segment, respectively. k is the number of parameters estimated. n_1 and n_2 are the number of observations in the first and second segments, respectively. After comparing the test statistic F to the critical values from the F-distribution, if F exceeds the critical value, I reject the null hypothesis H_0 and conclude that there is a structural break at the specified point. If a structural break is identified, I consider the time series to be stationary and assume that the ADF test was biased due to the nature of the time series. The structural break can be addressed by removing outliers (see the next section). If there is no structural break, the result of the ADF test is not biased, and another order of differentiation is applied to make the time series stationary.

3.2.3. Outliers' detection and removal

After ensuring the time series is stationary, and before building the ARIMA model, I remove the outliers of the time series. Outliers are single data points that goes far outside the median value of a time series (Enderlein, 1987). Removing it improves model accuracy and reduces the likelihood of including mismeasured values or values representing implausible events (Appaia and Palraj, 2023). To detect outliers, I use the interquartile range (IQR) criterion (Enderlein, 1987). The IQR criterion indicates that observations below $q_{0.25} - 1.5 \cdot IQR$ or above $q_{0.75} + 1.5 \cdot IQR$ are considered as outliers, where $q_{0.25}$ and $q_{0.75}$ correspond to 1st and 3rd quartile respectively, and IQR is the difference between the 3rd and 1st quartiles. If a data point is identified as an outlier, it is replaced with the median value of the time series. This method is widely used in the forecasting literature to eliminate the unwanted effects of outliers (Appaia and Palraj, 2023). Some outliers may also be the cause of a structural break

according to Chow's definition. Therefore, removing an outlier from a time series that has a structural break and replacing it with the median value could resolve the structural break issue and significantly improve forecast accuracy.

3.2.4. Estimation of ARIMA parameters

I can now build the ARIMA model. The ARIMA model contains three sections, which are the auto-regressive (AR), the integrated (I), and the moving-average (MA), represented below with "p", "d", and "q" parameters, respectively. Each letter gives the amount of AR, I, or MA parameters. There exists a specific combination of these parameters for each time series that produces an ARIMA model which best fits the data. The ARIMA model for a time series Y_t can be expressed as (Box and Jenkins, 1970):

$$\left(1 - \phi_1 L - \cdots - \phi_p L^p\right)(1 - L)^d Y_t = \left(1 + \theta_1 L + \cdots + \theta_q L^q\right)\epsilon_t$$

where L is the lag operator, defined as $L^k Y_t = Y_{t-k}$, ϕ_i (for i $= 1, \ldots,$ p) are the parameters of the autoregressive part, and θ_j (for $j = 1, \ldots, q$) are the parameters of the moving average part, and ϵ_t is the residual at time t. In order to determine the parameters of our ARIMA models for China and India's CO2 emissions, I process in a two-step method. The first step is to use Box-Jenkins approach. It allows to have a first intuition about the values of p and q by using the autocorrelation function and the partial auto-correlation function (Box and Jenkins, 1970). The auto-correlation function (ACF) helps in identifying the order of the MA component, with a sharp cut-off in ACF after lag q, suggesting an MA(q) model. Similarly, the partial auto-correlation function (PACF) helps in identifying the order of the AR component, where a sharp cutoff in PACF after lag p and a steady decrease in ACF indicates an AR(p) model. The integrated component will be automatically set to d=0 as I already now that the data does not need to be further differentiated, as it is stationary. As these intuitions for p and q are only based on what I observe, I select several possible values for p and q. The second step involves validating one of our initial guesses for p and q from Box-Jenkins approach. I manually calculate the values of two error-related information criteria for the plausible models identified and select the pair of parameters that minimizes these error criteria. The Akaike information criterion (Akaike, 1973) balances model fit and complexity. The Bayesian information criterion (Schwarz, 1978) is similar to the AIC but introduces a stricter penalty for models with more parameters:

$$AIC = 2k - 2\ln(L)\ ;\ BIC = k\ln(n) - 2\ln(L)$$

where k is the number of parameters in the model, n is the number of observations, an L is the likelihood of the model, computed as the product of the probability densities of the residuals.

3.2.5. Tests for white noise and validity of regression model

Once an ARIMA model is built for each country's CO2 emissions, I must analyze if the model produces residuals that are said to be "white noise", to make sure that the proposed model is valid for interpretation. This implies checking that they are normally distributed, homoscedastic, not autocorrelated, and do not contain any structural break (Moffat and Akpan, 2019). The last condition is checked using the Chow test explained above.

Autocorrelation indicates that the residuals are correlated, meaning that the model has not captured all the underlying patterns in the data. To detect autocorrelation, I use the Ljung-Box test, which has the following hypotheses (Ljung and Box, 1978):

- H_0: The residuals are independently distributed (no autocorrelation).
- H_1: The residuals are not independently distributed (autocorrelation present).

The test statistic is the following:

$$Q = n(n+2)\sum_{k=1}^{m}\frac{\widehat{\rho_k^2}}{n-k} \sim \chi_m^2$$

where n is the sample size, m is the number of lags being tested, $\widehat{\rho_k}$ is the sample autocorrelation at lag k. If the p-value of the test is bigger than 5%, I conclude that there is no autocorrelation in the residuals.

For checking the normality assumption, which ensures the validity of inference tests, I use the Jarque-Bera test. It has the following hypotheses (Jarque and Bera, 1981):

- H_0: The residuals are normally distributed.
- H_1: The residuals are not normally distributed.

The test statistic (JB) is calculated as:

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \sim \chi_2^2$$

where n is the sample size, S is the skewness[9] of the residuals, K is the kurtosis[10] of the residuals. After comparing the test statistic to the critical value from the chi-square distribution with 2 degrees of freedom, if it exceeds the critical value, I reject the null hypothesis and conclude that the residuals are not normally distributed.

The next step is to test the homoscedasticity of residuals, which means that the variance of the residuals is constant across observations. Heteroscedasticity can lead to inefficient estimates and biased inference. In order to test for homoscedasticity, I use the Goldfeld-Quandt test. This test checks for the presence of heteroscedasticity by comparing the variances of residuals in different parts of the data. It has the following hypothesis (Goldfeld and Quandt, 1965):

- H_0: The residuals have constant variance (homoscedasticity).
- H_1: The residuals have changing variance (heteroscedasticity).

The test statistic F is calculated as:

$$F = \frac{RSS_1/(n_1 - k)}{RSS_2/(n_2 - k)} \sim F_{(n_1-k),(n_2-k)}$$

where RSS_1 and RSS_2 are the residual sum of squares for the first and second group, respectively, and n_1 and n_2 are the number of observations for each group. Then, I compare the calculated F statistic to the critical value from the F-distribution. If the calculated F is less than or equal to the critical value, I do not reject the null hypothesis (H_0) and consider the residuals are homoscedastic.

If the residuals are said to be white noise, then the forecast will be reliable. If not, the forecast won't be completely reliable, but there is not much to do as an alternative to have white noise results. As it is not reasonable to modify the time series, the only possibility would be to change the parameters of the ARIMA model, but then the model wouldn't be optimal anymore.

[9] Skewness $= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{s}\right)^3$ where X_i are the observations, $\overline{X}$ is the sample mean, and s the standard deviation.

[10] Kurtosis $= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}}{s}\right)^4$ where X_i are the observations, $\overline{X}$ is the sample mean, and s the standard deviation.

3.2.6. Forecast with ARIMA model

The next step is to forecast the model. The memory of a forecasting model, which is the number of past periods used for forecasting, limits the maximum number of future periods that can be forecasted. For an ARIMA model, it is the amount of moving average (MA) parameters (Box and Jenkins, 1970). ARIMA models are generally used for short to medium-term forecasting and may not perform well for long-term predictions. For this model, the objective is to forecast fossil CO2 emissions of China and India up to 2025, if the model allows it. The general formula for forecasting with an ARIMA model for period t is:

$$\widehat{y_t} = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

where $\widehat{y_t}$ is the forecasted value at time t, μ is the mean of the series, ϕ_i are the coefficients for the autoregressive terms, θ_j are the coefficients for the moving average terms, and ϵ_t is a residual.

3.2.7. Forecasting accuracy

I can now evaluate the quality forecasts using the mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE) criteria. These metrics assess the accuracy of the forecasts.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}| \quad ; \quad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2} \quad ; \quad MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right|$$

where y_i and $\widehat{y_i}$ represent the values of the original and forecasted time series, respectively.

3.2.8. Discussion & limitations

After estimating separate ARIMA models for China and India, it is interesting to identify the patterns and short-term forecasting capabilities specific to each country. Comparing the ARIMA forecasts could provide insights into the differences in their emissions trajectories.

3.3. Methodology for Vector Error Correction Modeling

3.3.1. Data cleaning and importation

After the whole "world development indicators" dataset is collected for both countries, a shorter database with only relevant variables is built in an Excel file. It regards 33 and 35 variables for China and India respectively. Then, some variables must be modified for the regression. Indeed, some variables are related to the economy of a country and are defined in LCU (local currency units) or in American dollars. These variables are either defined in constant or current monetary units. It is necessary to use the constant measure when available, because it removes the effect of inflation or deflation that the country experienced. This measure then reflects the real evolution of an indicator, independently of economic growth or decline. When the economic variable is expressed in a current currency, I remove the effect of inflation using the variable called "inflation, GDP deflator (annual %)", which measures the increase or decrease of a country's economic size. I create a variable called "cumulative inflation" for both China and India, which represents the cumulative inflation since 1961. The formula is as follows, where I initialize the cumulative inflation of 1961 at 1, as it is the first year of the period studied:

$$Cumulative\ inflation_{\ year\ t} = cumulative\ inflation_{year\ t-1} \times (1 + inflation_{year\ t})$$

Then, for each country and each variable initially measured in current currency units, I use the cumulative inflation to create a new variable, where i is an index for each variable:

$$variable_{i,t}\ (in\ constant\ units) = \frac{variable_{i,t}\ (in\ current\ units)}{cumulative\ inflation_t}$$

3.3.2. Time series stationarity

In this section, I ensure that all 33 and 35 explanatory variables for China and India are stationary, and if they are not initially, I transform the time series using the same methodology as in the ARIMA part.

3.3.3. Selection of regression variables for VECM

Before selecting the variables for the VECM model, I do a simple linear regression for each variable to observe which ones have the highest R-squared values and correlation with CO2 emissions. Then, I use a 3-steps methodology to select the variables that will be included in the VECM to forecast CO2 emissions. It is important to note that I assess the causality

between variables without taking out the outliers from the time series. Indeed, what needs to be found at that time of the process is potential causation. Removing outliers is important to forecast time series, but it can impact negatively the direct causal relationship.

In the first step, I remove from the selection all the variables that do not have the same differentiation order than the CO2 emissions variable, as they cannot be used in the VECM. Indeed, cointegrated time series, such as the ones of a VECM, require that the variables are integrated of the same order (Engle and Granger, 1987).

Second of all, from the remaining variables, I select the ones that granger-cause the CO2 emissions time series. It tests wheter the past values of a independent variable, as X, contains usefull information for the forecasting of Y, other that the information already contained in the past realisations of Y. This methodology of selecting variables for a VECM model by using granger causality is described by Engle and Granger (1987). The Granger Causality test (Granger, 1969) has the following hypotheses:

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (no granger-causality)
- H_1 : any $\beta_p \neq 0$

assuming a lag length of p and that the $\beta's$ are coefficients in the following regression:

$$X_t = c_1 + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + a_t$$

If the p-value of the F-test is lower than the 5% threshold, I reject the null hyppothesis and conclude that the variable X granger cause variable Y, or the CO2 emissions in this case. The variables selected with this methodology constitute the first variable set that will be tested for the VECM.

Finally, I perform the Least Absolute Shrinkage and Selection Operator (LASSO) method (Tibshirani, 1996). Liang and Schienle (2019) studied this variables selection method for VECM modeling and qualify it as “efficient” and “consistent” for high-dimensional systems, as it is in this research. Bennedsen et al. (2021) also used LASSO method to forecast US CO2 emissions. Overall, LASSO enhances prediction accuracy and interpretability by producing simpler models that select only the most important predictors. Indeed, it adds a penalty equal to the absolute value of the magnitude of coefficients to the regression model. This constraint causes some coefficients of the regression to be exactly zero. I consider the following model:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon$$

where Y is the response variable, $X_j$are the predictor variables, β_j are the coefficients, and ε is the error term. Then, the LASSO method estimates the coefficients by minimizing the following objective function (Tibshirani, 1996):

$$min_{\beta_0,\beta} \ \frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where N is the number of observations and λ is the regularization parameter that controls the amount of shrinkage applied to the coefficients. A larger λ results in more coefficients being shrunk to zero. To choose the optimal λ, I minimize the mean squared error (MSE) on the validation set. Because it uses a cross-validation technique to minimize lambda, a different optimal lambda is found each time the code is run, and then different variables are selected with LASSO algorithm. Then, I choose, after running the optimization code 10 times, a maximum of 5 variables that have together the highest adjusted R-squared value. The choice of taking maximum 5 predictors is arbitrary and seems to be neither too small nor too large. It limits the high number of variables that could be selected with LASSO. The predictors selected using this methodology will constitute the second set of variables to build a VECM model.

Before validating both sets, I ensure that there is no multicollinearity among the predictors, so that the causation relationships between fossil CO2 emissions and predictors, as well as potential forecasts, are reliable. To do so, I compute the variance inflation factor (VIF) for each explanatory variable X(i) (Marquardt, 1970):

$$VIF(X_i) \ = \ \frac{1}{(1 - R_i^2)}$$

where Ri2 is the coefficient of determination obtained by regressing Xi on all the other predictor variables in the model. If the VIF is higher than 5, which is a common rule used in the literature (Shrestha, 2020), I consider that the variable is too correlated with other variables, and remove it from the set. Then, I compare the R-squared value of the new set with the other “LASSO” potential sets to still choose the most together-relevant predictors.

3.3.4. Outliers' removal

After the most together relevant variables for fitting CO2 emissions are selected, I remove the outliers for each variable, to prepare the forecasting set. This follows the same methodology as in the ARIMA part, explained above.

3.3.5. VECM creation

The Vector Error Correction Model (VECM) is an econometric model used to capture the long-run and short-run dynamics of cointegrated time series data (Engle and Granger, 1987). The following procedure is used for both countries and both variables set selected using granger causality and LASSO methods. The usual approach to create the VECM model is to use Johansen's procedure (Johansen, 1991) for testing whether at least one cointegration between variables exists, which would indicate that there is at least one long-run equilibriums among the variables. If there is a long-run equilibrium, then a Vector Error Correction Model (VECM) can be estimated.

The first step of the Johansen procedure consists in determining the optimal lag-length for the underlying Vector Autoregressive (VAR) model for our dataset, represented as follow:

$$Y_t = c + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \cdots + \Phi_p Y_{t-p} + \epsilon_t$$

where $Y_j\ (j = 1, \dots, t)$ are a vector of n endogenous variables, $\Phi_i (i = 1, \dots, p)$ are coefficient vectors, and ϵ_t is the error term. I choose the lag length that minimizes Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC), as in the Johansen procedure.

The second step is to find the optimal amount of cointegration relationships between the variables. To do so, I perform the "Maximum Eigenvalue Test" on the following Vector Error Correction Model (VECM):

$$\Delta Y_t = \Pi Y_{t-1} + \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \cdots + \Gamma_{p-1} \Delta Y_{t-(p-1)} + \epsilon_t$$

where Δ is the difference operator, ΠY_{t-1} is used to take into account the long-term equilibrium relationship, with $\Pi = \alpha\ \beta$ '. Also, α is the matrix for the adjustment coefficients, and β the matrix of cointegration vectors. Finally, $\Gamma_{i-1} \Delta Y_{t-(i-1)}$ captures the short-term dynamics for lagged differences of the endogenous variables.

I still have to perform the Maximum Eigenvalue Test, which will indicate the number of cointegration relationships. The test statistic is:

$$\lambda_{\max}(r, r+1) = -T \ln\left(1 - \hat{\lambda}_{r+1}\right)$$

where T is the number of observations and lambda the eigenvalue of the Π matrix. The null hypothesis (H_0) is that there are r cointegrating vectors against the alternative of r+1 (H_1). After comparing the maximum eigenvalue statistic with the critical values from the Johansen table, beginning with r = 0 and until the null hypothesis isn't rejected, I can determine the number of cointegrating vectors. At the end of the Johansen procedure, all the parameters I need to implement for the VECM model are known: the optimal lag length to use, as well as the appropriate number of cointegration relationships.

3.3.6. White noise residuals

After the model has been specified, I give interests to the residuals of the model. The methodology is similar to the one described above in the ARIMA model part, where I test either the residuals are white noise or not.

3.3.7. Forecasts

First of all, the number of periods ahead forecasted must be determined. It is decided that it would be 8 periods ahead, so that forecasted CO2 emissions can be compared with the targets of both China and India for 2030. To forecast multiple steps ahead with a VECM, the following process is repeated iteratively, using the predicted values of Y and ΔY from previous steps as inputs for subsequent forecasts. I first compute the forecast for the first difference, initializing with $h = 1$, where h is the number of forecast period ahead. i is the lag order:

$$\Delta Y_{T+h} = \Pi Y_{T+h-1} + \sum_{i=1}^{k-1} \Gamma_i \Delta Y_{T+h-i}$$

Then, we can update the value Y:

$$Y_{T+h} = Y_{T+h-1} + \Delta Y_{T+h}$$

3.3.8. Quality assessment of forecasts

After the time series are forecasted, I assess the quality of it using the same methodology as in the ARIMA model part, where I use RMSE, MAPE, and MSE criteria.

3.3.9. Discussion & comparison

First, I discuss the results of the simple linear regressions by analyzing the R-squared values for each predictor. After that, I analyze both China's and India's VECM forecasts, which incorporates long-term cointegration relationships and the short-term dynamics of the predictors. I also do a comparison between countries. This comparison provides insights into the differences in their emissions trajectories, the effectiveness of their environmental policies, and their respective economic growth patterns.

3.4. Comparisons between ARIMA and VECM models

Finally, I compare ARIMA and VECM forecasts by evaluating the strengths and limitations of each methodology in predicting CO2 emissions. This helps determine the most appropriate approach for different forecasting scenarios and highlight the added value of considering multiple interacting variables in emissions forecasting.

4. Results and Limitations

4.1. ARIMA results

4.1.1. Stationarity

4.1.1.1. China

After doing a first ADF test on the unmodified annual CO2 emissions, the p-value is higher than 5%. Consequently, I take the logarithm of the time series and differentiate it once. After doing the ADF test on the modified time series, I obtain a p-value of 0.01943, which reject the null hypothesis of non-stationarity. The yearly fossil CO2 emissions of China are integrated of order 1.

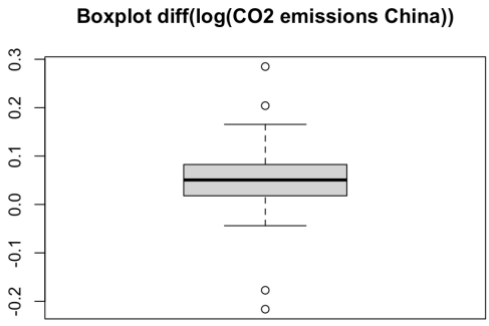
4.1.1.2. India

After doing a first ADF test on the unmodified annual CO2 emissions, the p-value is higher than 5%. Then, I take the logarithm of the time series and differentiate it once. After doing the ADF test on the modified time series, I obtain a p-value still higher than 5% (0.3834), which doesn't reject the null hypothesis of non-stationarity yet. Consequently, I apply Philipp-Peron test, and find out that the p-value of the test is lower than 1%, which is contradictory with ADF results. Then, I apply Chow-test on the time series. After observing the time series, I identify one potential breaking points at n=58. The p-value of the Chow test for n=58 is lower than 5%, meaning that there is a structural break (Appendix 2). Because of the nature of India's CO2 emissions time series, the ADF test is probably biased. Looking at both ACF and PACF (section 4.1.3.2) confirms that the time series is integrated of order 1.

4.1.2. Outliers' replacement

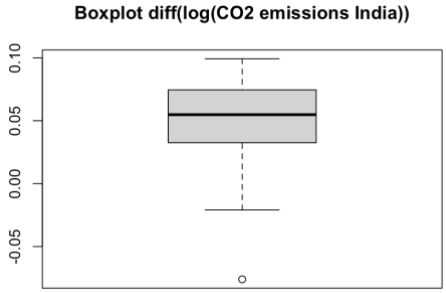
4.1.2.1. China

After plotting the boxplot on the transformed time series and computing interquartile range (IQR) criterion, I notice there are four outliers. I replace it by the median value of the time series. Below, each dot is an outlier:



4.1.2.2. India

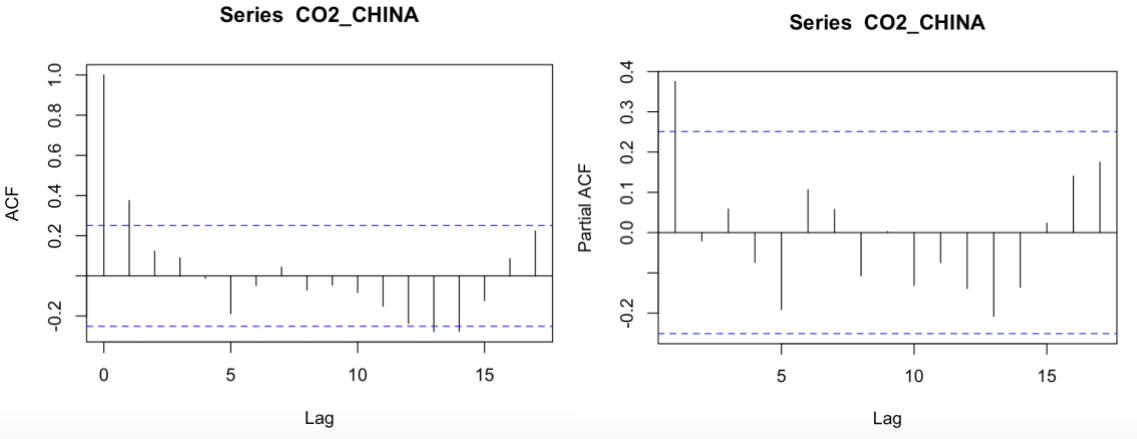
After plotting the boxplot on the transformed time series and computing interquartile range (IQR) criterion, there is one outlier at n=59, that I replace by the median value of the time series. Chow-test also proves that the replacement of this outlier removed the structural break located at n=58, which will make the forecast reliable.



4.1.3. Box and Jenkins approach and ARIMA model formation

4.1.3.1. China

I can now apply the Box and Jenkins approach to estimate the parameters of the best-fitting ARIMA model. The autocorrelation and partial autocorrelation functions are the following:

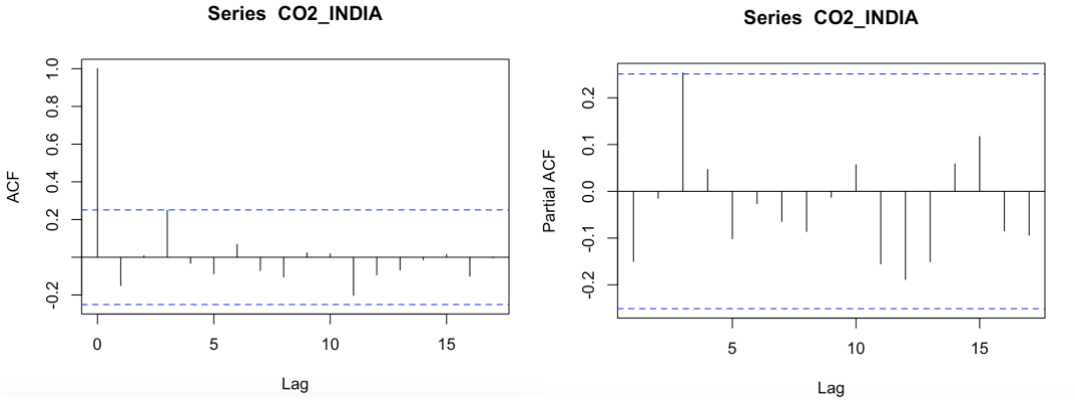


The PACF has 1 notorious spike for the first two lags, indicating an optimized auto-regressive parameter of probably 1. From the ACF, it seems that the moving-average component could be of 0 because it isn't sharply decaying. It could also be of 1 or 2, given the notorious spikes. After computing AIC and BIC criteria for possible ARIMA models, I found that the optimal ARIMA model is (1,0,0) for the difference of logarithms of fossil CO2 emissions:

$$\Delta \log(CO2\_emissions_t) = 0.0504 + 0.3768 \times \Delta \log(CO2\_emissions_{t-1}) + \epsilon_t$$

4.1.3.2. India

Applying the same methodology for India, here are the ACF and PACF:



Both ACF and PACF have one notorious spike at lag 3, but they are nearly out of the bounds. Then, both auto-regressive and moving-average parameters could be of 0 or 3. After computing AIC and BIC criteria for the four possible ARIMA models, the optimal model is ARIMA (0,0,3):

$$\Delta \log(\text{CO2\_emissions}_t) = 0.0545 - 0.1986\epsilon_{t-1} + 0.1717\epsilon_{t-2} + 0.4000\epsilon_{t-3} + \epsilon_t$$

4.1.4. White noise tests

4.1.4.1. China

After applying Ljung-box test to check the autocorrelation, Goldfeld-Quandt test to check the homoscedasticity, and Jarque-Bera test to check the normality, I notice that residuals are white noise:

```
        Ljung-Box test

data:  Residuals from ARIMA(1,0,0) with non-zero mean
Q* = 5.5592, df = 9, p-value = 0.7831

Model df: 1.   Total lags used: 10
```

```
        Jarque Bera Test

data:  arima_CHN$residuals
X-squared = 2.6554, df = 2, p-value = 0.2651
```

```
        Goldfeld-Quandt test

data:  residuals_AR_CHN
GQ = 0.88375, df1 = 30, df2 = 29, p-value = 0.6311
alternative hypothesis: variance increases from segment 1 to 2
```

There is no structural break in the data, as after using the Chow-test on potential breaking points outlined on the plot (n=14,35,45), the test statistics aren't significant (Appendix 3).

4.1.4.2. India

Using the same methodology, I notice that the ARIMA (0,0,3) residuals for India CO2 emissions are also white noise:

```
        Ljung-Box test

data:  Residuals from ARIMA(0,0,3) with non-zero mean
Q* = 2.4824, df = 7, p-value = 0.9284

Model df: 3.   Total lags used: 10
```

```
        Jarque Bera Test

data:  arima_IND$residuals
X-squared = 1.3139, df = 2, p-value = 0.5184
```

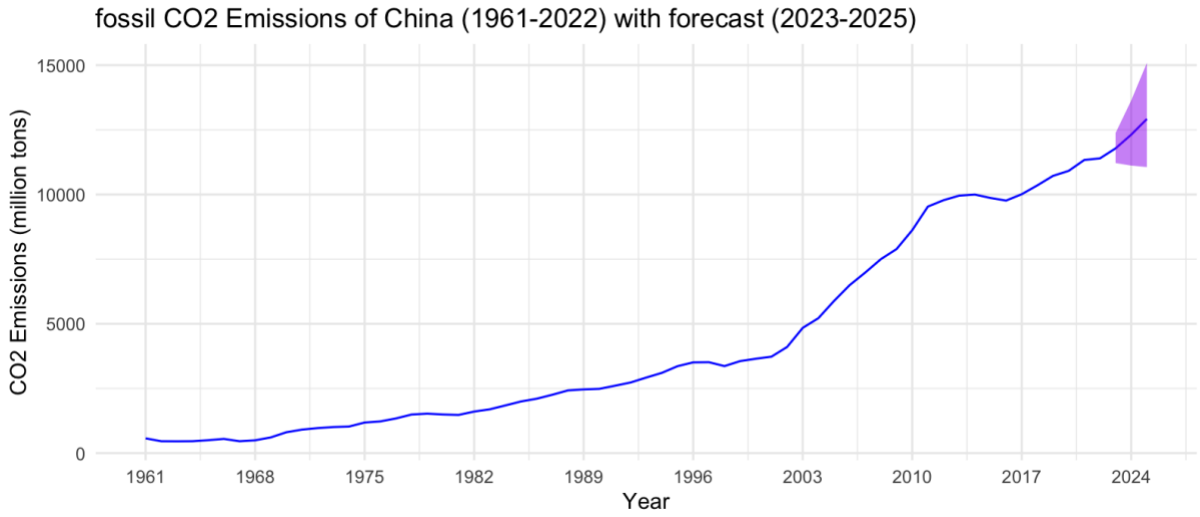
```
                Goldfeld-Quandt test

data:  residuals_AR_IND
GQ = 0.76619, df1 = 30, df2 = 29, p-value = 0.7638
alternative hypothesis: variance increases from segment 1 to 2
```

There is no structural break in the data, as I already checked it in section 4.1.1.2.

4.1.5. Forecast

4.1.5.1. China

Because our model has one AR component, I can forecast the CO2 emissions of China up to 2025, as wanted. Using the iterative formula presented in the methodology, the CO2 emissions of China are expected to increase up to 2025. All three forecasted values (differences of logarithms) are above zero. The pink zone represents the 80% confidence interval for the forecasted three years ahead:

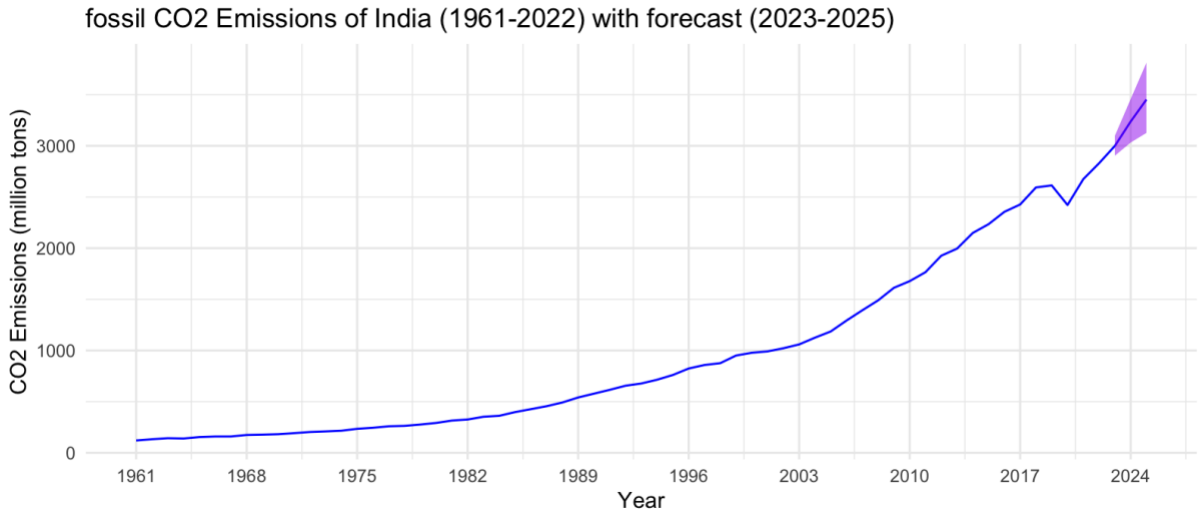


Knowing that the carbon dioxide emissions of China in 2022 were of 11396.78 million tons and that the three forecasted values are of 0.0334, 0.0440, and 0.0479 for 2023, 2024, and

2025 respectively, this gives an estimated CO2 emissions level of 12879.9 million tons for 2025.

4.1.5.2. India

Because our model has three MA component, I can forecast the CO2 emissions of India up to 3 periods ahead maximum, or 2025. The CO2 emissions of India are expected to increase up to 2025. All three forecasted values (differences of logarithms) are above zero. The pink zone represents the 80% confidence interval for the forecasted three years ahead:



Knowing that the carbon dioxide emissions of India in 2022 were of 2829.64 million tons and that the three forecasted values are of 0.0580, 0.0764, and 0.0642 for 2023, 2024, and 2025 respectively, this gives an estimated CO2 emissions level of 3456.13 million tons for 2025.

4.1.6. Quality forecast assessment

4.1.6.1. China

After computing three error-related criteria to assess the forecasting quality, I find a MAPE value of 157.06%, a RMSE value of 0.0378, and a MASE value of 0.871 (Appendix 4).

4.1.6.2. India

Similarly, for India fossil CO2 emissions, I find a MAPE value of 100.32%, a RMSE value of 0.0245, and a MASE value of 0.631 (Appendix 5). We can notice that the ARIMA model for India performed better, with all three error-related criteria being lower. The MAPE measure

shows that it is 56,7% more accurate. This makes sense regarding the softer evolution of Indian fossil CO2 emissions over time, which is better forecasted with an ARIMA model.

4.1.7 Discussion and comparison between China and India

At this stage of the analysis, there is not many insights to highlight, due to the univariate nature of ARIMA modeling. Nevertheless, it is interesting to compare the expected growth evolution of emissions between both countries. The CO2 emissions of India are expected to increase much faster than Chinese emissions. This makes sense with the recent evolution of fossil CO2 emissions of both countries.

4.2. VECM results

4.2.1. Stationarity of explanatory variables

4.2.1.1. China

I check for stationarity using the same methodology as in ARIMA part, taking logarithms, and using Augmented Dickey-Fuller test first, and then Philipp Peron test if necessary. It results that 10 out of the 33 variables are I(2) or I(0) and then cannot be considered for the VECM. Results are available in the R-Studio file.

4.2.1.2. India

Similarly, for India's world development indicators, 9 out of the 35 variables are I(2) or I(0) and then cannot be considered for the VECM. Results are available in the R-Studio file.

4.2.2. Variables selection

4.2.2.1. China

Firstly, after using Granger-causality test, there is only the variable "Gross national expenditure (constant LCU)" that Granger cause China's fossil CO2 emissions. Then, I built the second set of predictors using the LASSO method. After running the code 10 times, 3 different outputs arise (Appendix 6). After comparing both adjusted R-squared value in a linear regression, the best "LASSO" set is composed of 4 variables: agricultural land (squared meters), value-added of industry (constant LCU), Gross capital formation (constant LCU), and Exports of merchandise (constant US dollars, adjusted) (Appendix 7), with an adjusted r-squarred value of 0.5832 in a linear regression. Then, I ensure that there isn't a high degree of multicollinearity among the selected predictors using variance inflation factor (VIF) test. It appears that the level of multicollinearity between every variable is very small (Appendix 8).

4.2.2.2. India

Using the same methodology as for China, it appears that out of the 26 potential predictors, not one Granger cause the fossil CO2 emissions of India. This first methodology does not help to select variables for VECM model in this case. After applying LASSO method on the 26-variables set, 3 different outputs arise (Appendix 9). Out of the 3 predictors set, one has more than five variables. For this set, I select the five variables with the most significant coefficients in the regression. Finally, after comparing the three sets, the most together significant variables are: agricultural land (squared meters), value-added of industry (constant LCU), final consumption expenditure (constant LCU), gross domestic savings (% of GDP), and DEC alternative conversion factor (LCU per US$), with an adjusted R-squared value of 0.3616 in a linear regression (Appendix 10). After using VIF test, it appears that the level of multicollinearity among the LASSO selected predictors is very small (Appendix 11).

4.2.3. Outliers' replacement

4.2.3.1. China

Some outliers were present in all four variables. They are replaced by the median value of the time-series, for consistency in methodology with the CO2 emissions part.

4.2.3.2. India

Some outliers were present in all five variables. They are replaced by the median value of the time-series, for consistency in methodology with the CO2 emissions part.

4.2.4. VECM creation

4.2.4.1. China

We have already seen earlier in the methodology that all the time series must have the same stochastic trend properties. There are two potential sets to fit a VECM model: the variable selected with Granger-causality, and the variables selected with LASSO method. All predictors of both sets and China's fossil CO2 emissions are I(1). I can regress one time series on another, according to Phillips (1986).

First, it is important here to choose for the right lag length in our model. A poor lag length won't enable us to estimate precisely the parameters and if it is too long, some degrees of freedom will be lost. I choose a lag length that is sufficient for the residuals to be white noise. I then use the Akaike Information Criteria to determine the optimal lag length. It turns out that for both "Ganger" and "LASSO" datasets, it is of 1.

I now carry out the Johansen procedure. For the "Granger" dataset, we can see after using Johansen procedure that there is no cointegration relationship (Appendix 12). I can't estimate a VECM model only with the variable that Granger cause China's fossil CO2 emissions, and I pursue the process with the LASSO-selected variables. For the LASSO set, the test statistics is higher than the critical value at 1% significance level for r = 0. This means that I reject the null hypothesis that: r = 0. But the test statistic at r ≤ 1 is lower than the critical value. I do not reject H_0 and choose r = 1 (Appendix 13). I conclude that there is 1 cointegration relationship for the LASSO variables.

I can now estimate the VECM equations (Appendix 14). The equation for CO2 emissions is given below:

$$\begin{aligned}\Delta CO2_t = &-0.0820.(\mathrm{ECT}_{t-1}) - 0.0094 - 0.1877.\Delta CO2_{t-1} + 0.2215.\Delta AGRI_{t-1} \\ &- 0.0779.\Delta INDUSTRY_{t-1} - 0.0469.\Delta CAP\_FORM_{t-1} \\ &- 0.0009.\Delta MERCH\_EXP_{t-1} + \epsilon_t\end{aligned}$$

4.2.4.2. India

Similarly, because both the variables selected from LASSO methodology and India's fossil CO2 emissions are I(1), I can build a VECM model. Akaike Information Criteria determined that the optimal lag length is of 1. Using Johansen procedure, I reject sequentially the null hypotheses that: r = 0, r <= 1, and r <= 2. I can't reject r <= 3 and conclude that there are 3 cointegration relationships using 1% significance level (Appendix 15). I can now estimate the VECM equations (Appendix 16). The equation for CO2 emissions is given below:

$$\begin{aligned}\Delta CO2_t = &-0.8839.ECT1_{t-1} + 0.2172.ECT2_{t-1} + 0.2846.ECT3_{t-1} + 0.0165 \\ &- 0.1222.\Delta CO2_{t-1} - 0.0735.\Delta AGRI_{t-1} - 0.2047.\Delta FINAL\_CONS_{t-1} \\ &- 0.0113.\Delta INDUSTRY_{t-1} - 0.1166.\Delta DOM\_SAVINGS_{t-1} \\ &- 0.1027.\Delta DEC\_CHG_{t-1} + \epsilon_t\end{aligned}$$

4.2.5. White noise tests

4.2.5.1. China

Using several Rstudio functions to test for autocorrelation (Ljung-box), homoscedasticity (Goldfeld-Quandt), and normality (Jarque-Bera), the residuals are white noise:

```
        Ljung-Box test

data:  Residuals
Q* = 14.945, df = 10, p-value = 0.1341

Model df: 0.   Total lags used: 10
```

```
        Jarque Bera Test

data:  resi_vecm_chn
X-squared = 0.40688, df = 2, p-value = 0.8159
```

```
        Goldfeld-Quandt test

data:  residuals_tvecm_CHN
GQ = 0.78288, df1 = 29, df2 = 28, p-value = 0.742
alternative hypothesis: variance increases from segment 1 to 2
```

4.2.5.2. India

Similarly, the residuals of VECM for India's fossil CO2 emissions are white noise:

```
        Ljung-Box test

data:  Residuals
Q* = 9.4598, df = 10, p-value = 0.4891

Model df: 0.   Total lags used: 10
```

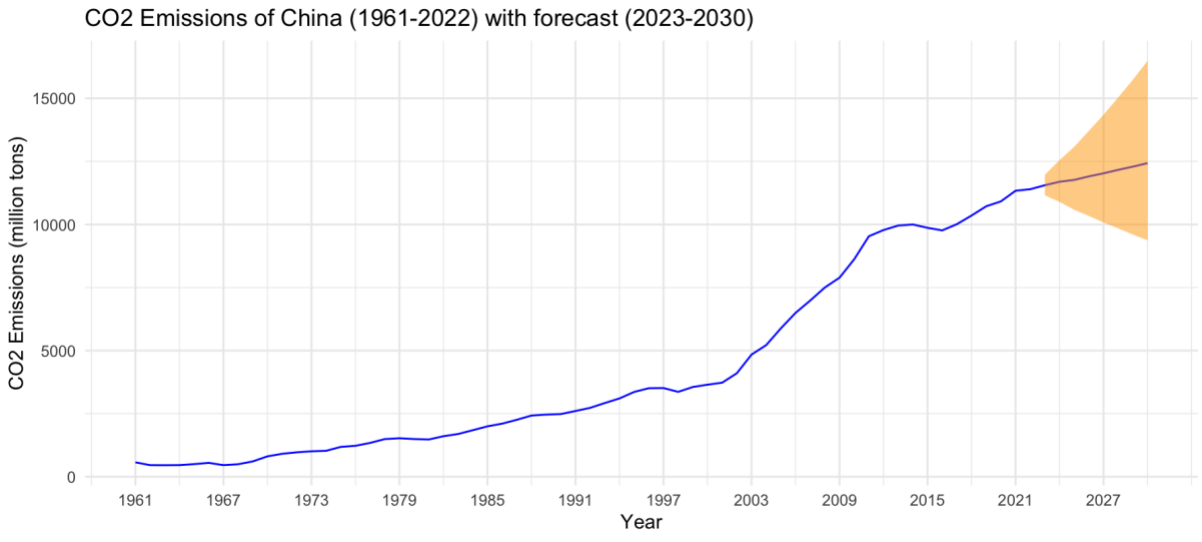
```
        Jarque Bera Test

data:  resi_vecm_ind
X-squared = 1.2138, df = 2, p-value = 0.545
```

```
        Goldfeld-Quandt test

data:  residuals_tvecm_IND
GQ = 0.96607, df1 = 29, df2 = 28, p-value = 0.5373
alternative hypothesis: variance increases from segment 1 to 2
```

4.2.6. Forecast

4.2.6.1. China

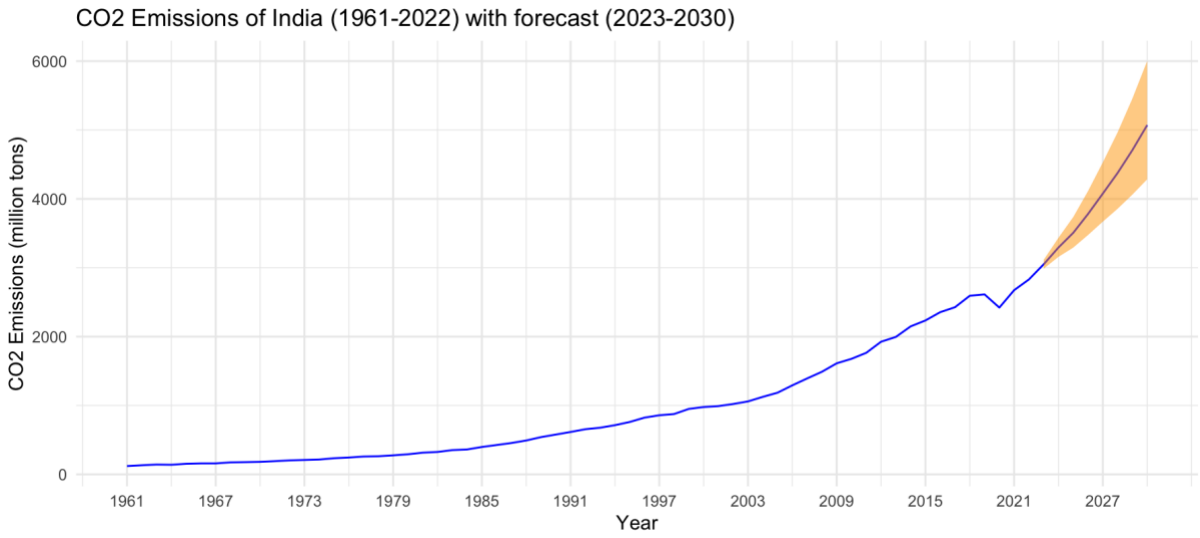
Using the iterative formula presented in the methodology and VECM model developed, the CO2 emissions of China are expected to increase up to 2030. The orange zone represents the 80% confidence interval for the forecasted eight years ahead:



Knowing that the carbon dioxide emissions of China in 2022 were of 11396.78 million tons and that the three forecasted values are of 0.0305, 0.0176, 0.0191, 0.0226, 0.0214, 0.0232, 0.0225, 0.0239 between 2023 and 2030 respectively, this gives an estimated CO2 emissions level of 12293.2 million tons for 2025 and 13766.0 million tons for 2030.

4.2.6.2. India

Similarly, the CO2 emissions of India are expected to increase up to 2030. All eight forecasted values are above zero. The orange zone represents the 80% confidence interval for the forecasted eight years ahead:



Knowing that the carbon dioxide emissions of India in 2022 were of 2829.64 million tons and that the forecasted values, which are differences of logarithms, are of 0.0750, 0.0795, 0.0768, 0.0628, 0.0751, 0.0753, 0.0708, 0.0730, and 0.0742 between 2023 and 2030, this gives an estimated CO2 emissions level of 5100.96 million tons for 2030.

4.2.7. Quality forecast assessment

4.2.7.1. China

After computing different error-related criteria, I find that the MAPE value is of 104.19, the MAE of 0.0304, and the RMSE of 0.0378 (Appendix 17). The VECM model, in average, made forecasts that are more precise than the ARIMA model, based on bot criteria.

4.2.7.2. India

Similarly, I find that the MAPE value is of 126.4, the MAE of 0.0540 and the RMSE of 0.062 (Appendix 18). In this case, the VECM model was less accurate than the ARIMA model.

4.2.8. Discussion

4.2.8.1. China

First, I discuss the results of the univariate linear regressions between the difference in logarithms of CO2 emissions and the 33 variables studied (Appendix 1). This gives an idea of how well an independent variable can explain the variation in the dependent variable. Out of the 33 variables, the predictors related to the economic activity of China performed the best, such as the gross domestic product, the gross national income, or the gross capital formation. Predictors related to production also performed well, with "industrial and construction added-value" having the highest R-squared value among all the predictors. The variable "aquaculture production" was also significant, with a p-value of 7%. Interestingly, the variable "fertilizer consumption (kilograms per hectare of arable land)" was also significant to explain the variation in fossil CO2 emissions.

It is now interesting to compare the forecast made for 2030 and the targets in CO2 emissions. The forecast made using VECM model only represents one possible scenario, where fossil CO2 emissions evolve at a rhythm implied by the recent evolution of CO2 emissions and the four variables it is cointegrated with, such as the added value of industry. This been considered for the discussion, it is now interesting to compare the forecast made for 2030 and the targets in CO2 emissions. In 2021, the government of China announced, "carbon dioxide emissions per unit of GDP will have dropped by more than 65% compared with the 2005 level, successfully achieving carbon dioxide peaking before 2030" (State Council of China, 2021). First, given the forecast made, China isn't expected to see its CO2 emissions peaking before 2030 because its fossil carbon dioxide emissions, which represents on average 98.74% of China's CO2 emissions (average between 2018 and 2022) (Global Carbon Atlas and GCP, 2024[11]), are expected to increase until 2030 at least if the country continues to keep the same economic and production growth as in the recent years. Regarding the second objective, we can compare the carbon intensity (CI)[12] of 2005 and the forecasts of the same indicator for

[11] The Global Carbon Project doesn't include CO2 emissions from land use change in its CO2 emissions estimations, but they post it separately. Because it is the only main item missing in annual CO2 emissions computations, one can estimate the share of GCP's CO2 emissions (fossil, cement, and lime production) of total CO2 emissions.

[12] Carbon intensity is defined as Kg of CO2 emitted per unit of real GDP (SOURCE).

2030. Using the data already collected and forecast of CO2 emissions, emissions are expected to increase by a factor of 2.34 between 2005 and 2030. For China's CI to decrease of at least 65%, the real GDP of China in 2030 will have to be at least 6.68 times larger than the one of 2005 if we consider the VECM forecast. Using S&P forecast of GDP increase of an average of 4.4% for the period 2022-2030 (Kuijs, 2022), the growth in real GDP would be of 5.21, which is below the one necessary to reach target.

Using the VECM model, it would have been interesting to explore the situations in which China would decrease its CI indicator by 2.85 between 2005 and 2030, instead of the 2.19 to 2.23 estimates. Unfortunately, as the model built to forecast CO2 emissions integrate variables that are related to the economic growth and that these variables are not expected to decrease in the next years (result from both VECM forecast and public information), it wouldn't be relevant to analyze to which extent industrial production or goods' exports should decrease to reach China's carbon intensity targets. Consequently, the only set of solutions of China to reduce CI is either to burn less fossil fuel and opting for more green energy sources such as solar or nuclear, or by using machinery that use fossil fuel but in a more efficient way in comparison to 2005, or by a combination of both techniques.

4.2.8.2. India

Similarly, as for China, variables related to economic growth and production have the highest R-squared values. In this category, there are for instance gross national expenses, added-value of industry, or added-value of services. It is also noticeable that, on average, India's CO2 emissions were less correlated to the 33 predictors than China's CO2 emissions. The variables "inflation in consumer prices" and "export as a capacity to import" that Bennedsen et al. (2021) have found to be highly correlated to US CO2 emissions, didn't have high r-squared values for India's analysis, with 5.4% and 1.1% respectively.

In the second part of this discussion, I compare the forecast made for 2030 and India's government target for their CO2 emissions level. In 2022, the government announced, "India updated its NDC according to which target to reduce emissions intensity of its GDP has been enhanced to 45 percent by 2030 from 2005 level" (India government, 2022). Using the forecast made with VECM model, the fossil CO2 emissions, which represents on average 99.5% of India's CO2 emissions (average between 2018 and 2022) (Global Carbon Atlas and GCP, 2024) would increase by a factor of 4.30 between 2005 and 2030. For India's CI to decrease of at least 45%, the real GDP of India in 2030 will have to be at least 9.56 times

larger than the one of 2005 if we consider the VECM forecast. If we extrapolate the growth in real GDP that happened between the period 2005-2022 to the period 2023-2030, which is of 6.03%, this would give an overall increase of the GDP by a factor of 4,32. Even if I use S&P's forecast of nominal GDP increase, which averages 9.6% for the period 2022-2030 (Biswas, 2023), the same factor would be of 5.64, which is also below the one necessary to reach target. Overall, the CI indicator of India is expected to decrease by a factor between 1.01 and 1.31 for the 2005-2030 period, instead of the 2.22 target.

As for China analyses, it wouldn't be relevant to analyze to which extent economic activity or industrial production should decrease to reach government's targets, given the expected increase in GDP from S&P (Biswas, 2023), and the forecasted evolution of production variables using VECM statistical model. The country should use cleaner energy sources, burn fewer fossil fuels by using new efficient technologies, or a combination of both methods.

4.3. Comparison between ARIMA and VECM models

In this section, the forecasting quality of both ARIMA and VECM models is discussed. Initially, the purpose of estimating an ARIMA model for each country was then to compare its performance with a more advanced VECM model, which incorporates predictors. For China, the VECM was more accurate than the ARIMA with a mean absolute percentage error (MAPE) value 52.9% lower. The root mean squared error (RMSE) and mean absolute error (MAE) showed that the ARIMA model was more accurate. Although the ARIMA model has lower absolute errors on average compared to the VECM model, the VECM model's predictions are relatively closer to the actual values when considering the size of the values. For India, all metrics showed that the ARIMA model was more accurate than the VECM model. The MAPE is 15.3% smaller for the ARIMA model than the VECM model. It seems that the VECM model only performed better than the ARIMA model for the analysis of China's fossil CO2 emissions, with a lower percentage error. The fact that it is the opposite for India could make sense, if we look at the evolution of the time series. Indeed, the evolution of India's CO2 emissions through time follows a very "smooth trend", which then doesn't require many predictors to foresee the time series, in contrary to China, where CO2 emissions are sometimes subject to short and important changes. These sudden changes cannot be explained well with past information only, such as with an ARIMA model, but are well explained using a VECM model that incorporates several predictors.

4.4. Limitations

This master's thesis, which forecasts fossil CO2 emissions for India and China using VECM and ARIMA models, presents several limitations. First, the data has some disadvantages. It covers a sixty-two-year period, from 1961 to 2022, which is maybe too long to reflect the recent changes in the relationships between CO2 emissions and predictors. Additionally, the sources and estimation methods of the data may be inaccurate, and the number of relevant predictors (33 variables) is significantly fewer than those available for the USA as shown by the McCracken and Ng (2020) database, for instance. Secondly, the models themselves have limitations. While VECM incorporates explanatory variables, potentially improving its accuracy, the model's parameters partially rely on historical data and relationships since 1961, which might not hold in the future. It has also been showed that this model didn't always perform better that the ARIMA model, though it doesn't account for external predictors. Finally, the approach used in this paper doesn't have the capability to incorporate "new" information and to predict changes in economic activity and energy production methods, which could significantly impact CO2 emissions in the coming years. Consequently, while the forecasts provide valuable insights, they only represent the situation where economic and production growths evolve at a rhythm based on what happened in the recent years. Finally, because the variability of fossil CO2 emissions isn't fully explained with the predictors selected, the forecasts miss some explanatory information that could be revealed by incorporating new economic, production, or other relevant predictors.

5. Conclusion

In this research paper, I first focused on studying the relations between different macro-economic indicators and CO2 emissions. The analyses confirm the findings of Bennedsen et al. (2021), demonstrating that variables related to economic activity and production are the most effective in explaining variations in CO2 emissions.

The forecasts indicate China is unlikely to peak CO2 emissions before 2030 if current growth rates continue. To achieve a CI reduction of 65%, and if China's fossil CO2 emissions will increase by a forecasted factor of 2.34, China's GDP would need to be 6.68 times larger than in 2005. The S&P analyses suggest expected growth rates will fall short. For India, economic growth and production-related variables showed strong but lower correlations with CO2 emissions compared to China. India's target is to reduce emissions intensity of its GDP by 45% by 2030 from the 2005 level. The VECM model forecast indicates India's fossil CO2 emissions will increase by a factor of 4.30 between 2005 and 2030. To meet the CI reduction target, India's GDP would need to increase by a factor of 9.56, but current growth projections suggest it will fall short too.

For both China and India, the only viable option to improve carbon intensity (CI) is to invest in more carbon-efficient technologies and produce cleaner energy, as reducing economic activity and production is not viable. They should focus on advancing renewable energy sources like solar, wind, and hydropower, as well as implementing energy-efficient technologies such as advanced manufacturing processes, smart grids, and carbon capture to meet their CI reduction targets (Chateau, 2023) (National Development and Reform Commission (NDRC) People's Republic of China, 2021).

As for later, a possible next step would be to include variables that are indirectly or directly relating to energy efficiency and to see their impact on the forecast of CO2 emissions. This would allow to do scenario analysis, and maybe give better forecasting accuracy. Unfortunately, such variables aren't available, or at least publicly. Using the forecasts made in this paper, future research could focus on developing strategies for China and India to bridge the gap between the forecasted carbon intensity and the established targets. For example, a combination of green energy adoption and advancements in fuel technologies could be explored to achieve these goals.

6. Bibliography

Abdullah, L., & Pauzi, H. M. (2015). METHODS IN FORECASTING CARBON DIOXIDE EMISSIONS : A DECADE REVIEW. Jurnal Teknologi, 75(1). https://doi.org/10.11113/jt.v75.2603

Aboul Ela, H. A. (2023). "Forecasting Emissions of Carbon Dioxide (CO2), Methane (CH4) and Energy Consumption in Egypt Using VECM and ARIMAX Models," Journal of Statistics Applications & Probability: Vol. 12: Iss. 3, Article 6. DOI: http://dx.doi.org/10.18576/jsap/120306

About | Climate Watch. Online https://www.climatewatchdata.org/about/faq/ghg, consulted on 17 April 2024

Aftab, S., Ahmed, A., Chandio, A. A., Korankye, B. A., Ali, A., & Fang, W. (2021). Modeling the nexus between carbon emissions, energy consumption, and economic progress in Pakistan : Evidence from cointegration and causality analysis. Energy Reports, 7, 4642-4658. https://doi.org/10.1016/j.egyr.2021.07.020

Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle. In: Petrov, B.N. and Csaki, F., Eds., International Symposium on Information Theory, 267-281.

Algieri, B., Iania, L., & Leccadito, A. (2023). Looking ahead : Forecasting total energy carbon dioxide emissions. Cleaner Environmental Systems, 9, 100112. https://doi.org/10.1016/j.cesys.2023.100112

Andrew, R., & Peters, G. P. (2022). The Global Carbon Project's fossil CO2 emissions dataset: 2022 release. CICERO Center for International Climate Research, Oslo, Norway.

Appaia, L., & Palraj, S. (2023). On replacement of outliers and missing values in time series. EQA - International Journal of Environmental Quality, 53(1), 1–10. https://doi.org/10.6092/issn.2281-4485/16184

Bazaz, A. B., & Seksaria, V. (2014). (De)Constructing India's and China's Energy and Climate Change Policy Choices : Interfaces of Co-operation. Procedia - Social And Behavioral Sciences, 157, 322-329. https://doi.org/10.1016/j.sbspro.2014.11.035

Bennedsen, M., Hillebrand, E., & Koopman, S. J. (2021). Modeling, forecasting, and nowcasting U.S. CO2 emissions using many macroeconomic predictors. Energy Economics, 96, 105118. https://doi.org/10.1016/j.eneco.2021.105118

Biswas, R. (2023, 8 décembre). India seizes crown of fastest growing G20 economy. IHS Markit. https://www.spglobal.com/marketintelligence/en/mi/research-analysis/india-seizes-crown-of-fastest-growing-g20-economy-dec23.html

Bokde, N. D., Tranberg, B., & Andresen, G. B. (2021). Short-term CO2 emissions forecasting based on decomposition approaches and its impact on electricity market scheduling. Applied Energy, 281, 116061. https://doi.org/10.1016/j.apenergy.2020.116061

Boontome, P., Therdyothin, A., & Chontanawat, J. (2019). FORECASTING CARBON DIOXIDE EMISSION AND SUSTAINABLE ECONOMY : EVIDENCE AND POLICY RESPONSES. International Journal Of Energy Economics And Policy, 9(5), 55-62. https://doi.org/10.32479/ijeep.7918

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2008). Time Series analysis. Wiley series in probability and statistics. https://doi.org/10.1002/9781118619193

Chateau, J. (2023). A Framework for Climate Change Mitigation in India. IMF Working Paper, 2023(218), 1. https://doi.org/10.5089/9798400247002.001

China - Countries & Regions - IEA. (s. d.). IEA. Online https://www.iea.org/countries/china/emissions, consulted on 5 June 2024

Chow, G. C. (1960). Tests of Equality Between Sets of Coefficients in Two Linear Regressions. Econometrica, 28(3), 591–605. https://doi.org/10.2307/1910133

EDGAR - The Emissions Database for Global Atmospheric Research. (s. d.). Online https://edgar.jrc.ec.europa.eu/report_2023, consulted on 22 February 2024

Enderlein, G. 1987. “Hawkins, DM: Identification of Outliers. Chapman and Hall, London–New York 1980, Biometrical Journal 29 (2): 198–98.

Engle, R. F., & Granger, C. W. J. (1987). Co-Integration and Error Correction: Representation, Estimation, and Testing. Econometrica, 55(2), 251–276. https://doi.org/10.2307/1913236

Exploring the social challenges of low-carbon energy policies in Europe. (s. d.). European Environment Agency. Online https://www.eea.europa.eu/publications/exploring-the-social-challenges-of, consulted on 29 June 2024

Faruque, M. O., Rabby, M. A. J., Hossain, M. A., Islam, M. R., Rashid, M. M. U., & Muyeen, S. (2022). A comparative analysis to forecast carbon dioxide emissions. Energy Reports, 8, 8046-8060. https://doi.org/10.1016/j.egyr.2022.06.025

Fatima, S., Ali, S. S., Zia, S. S., Hussain, E., Fraz, T. R., & Khan, M. S. (2019). Forecasting Carbon Dioxide Emission of Asian Countries Using ARIMA and Simple Exponential Smoothing Models. International Journal Of Economic And Environmental Geology, 10(1), 64-69. https://doi.org/10.46660/ojs.v10i1.219

Friedlingstein, P., O’Sullivan, M., Jones, M., Andrew, R. M., Gregor, L., Hauck, J., Quéré, C. L., Luijkx, I., Olsen, A., Peters, G., Peters, W., Pongratz, J., Schwingshackl, C., Sitch,

S., Canadell, J., Ciais, P., Jackson, R., Alin, S., Alkama, R.,. . . Zheng, B. (2022). Global Carbon Budget 2022. Earth System Science Data, 14(11), 4811-4900. https://doi.org/10.5194/essd-14-4811-2022

Global Carbon Atlas. (2023, 5 décembre). Carbon Emissions - Global Carbon Atlas. Online https://globalcarbonatlas.org/emissions/carbon-emissions/, consulted on 5 March 2024

Global Carbon Project (GCP). (s. d.). About GCP. Online https://www.globalcarbonproject.org/about/index.htm, consulted on 16 June 2024

Goldfeld, S.M. and Quandt, R.E. (1965) Some Tests for Heteroscedasticity. Journal of the American Statistical Association, 60, 539-547. https://doi.org/10.1080/01621459.1965.10480811

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". Econometrica. 37 (3): 424–438. doi:10.2307/1912791.

Gütschow, J., & Pflüger, M. (2023). The PRIMAP-hist national historical emissions time series (1750-2022) v2.5 [Base de données]. Dans Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.10006301

Hyndman, R.J.; Athanasopoulos, G. Forecasting: Principles and Practice. Available online: www.OTexts. com/fpp (accessed on 5 September 2016).

Hyndman, R.J.; Khandakar, Y. Automatic Time Series Forecasting: The forecast Package for R. J. Stat. Softw. 2008, 27, 1–22.

India - Countries & Regions - IEA. (s. d.). IEA. Online https://www.iea.org/countries/india/emissions, consulted on 18 July 2024

India achieves two targets of Nationally Determined Contribution well ahead of the time. (s. d.). https://pib.gov.in/PressReleasePage.aspx?PRID=1987752#:~:text=In%20August%202022%2C%20India%20updated,enhanced%20to%2050%25%20by%202030, consulted on 26 July 2024

Jarque, Carlos M.; Bera, Anil K. (1981). "Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence". Economics Letters. 7 (4): 313–318. doi:10.1016/0165-1765(81)90035-5

Jena, P. R., Managi, S., & Majhi, B. (2021). Forecasting the CO2 Emissions at the Global Level : A Multilayer Artificial Neural Network Modelling. Energies, 14(19), 6336. https://doi.org/10.3390/en14196336

Johansen, Soren (1991). "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models". Econometrica. 59 (6): 1551–1580. doi:10.2307/2938278

Jones, M. W., Peters, G. P., Gasser, T., Andrew, R. M., Schwingshackl, C., Gütschow, J., Houghton, R. A., Friedlingstein, P., Pongratz, J., & Quéré, C. L. (2023). National contributions to climate change due to historical emissions of carbon dioxide, methane, and nitrous oxide since 1850. Scientific Data, 10(1). https://doi.org/10.1038/s41597-023-02041-1

Kohzadi, N., Boyd, M. S., Kaastra, I., Kermanshahi, B. S., & Scuse, D. (1995). Neural Networks for Forecasting : An Introduction. Canadian Journal Of Agricultural Economics/Revue Canadienne D Agroeconomie, 43(3), 463-474. https://doi.org/10.1111/j.1744-7976.1995.tb00135.x

Köne, A. Ç., & Büke, T. (2010). Forecasting of CO2 emissions from fuel combustion using trend analysis. Renewable And Sustainable Energy Reviews, 14(9), 2906-2915. https://doi.org/10.1016/j.rser.2010.06.006

Kuijs, L. (2022). Economic Research : China's trend growth to slow even as catchup continues. S&P Global Ratings. Online https://www.spglobal.com/ratings/en/research/articles/221110-economic-research-china-s-trend-growth-to-slow-even-as-catchup-continues-12550713

Kumari, P., Mishra, G.C., Pant, A.K., Shukla, G. and Kujur, S.N. (2014). Autoregressive Integrated Moving Average (ARIMA) Approach for Prediction of Rice (Oryza Sativa l.) Yield in India. The Bioscan-An International Quarterly Journal of Life Science, 9, 1063-1066.

Liang, C., & Schienle, M. (2019). Determination of vector error correction models in high dimensions. Journal Of Econometrics, 208(2), 418-441. https://doi.org/10.1016/j.jeconom.2018.09.018

Ljung, G.M. and Box, G.P. 1978. "On a Measure of a Lack of Fit in Time Series Models", Biometrika, 65.2, 297–303.

Mahan, M., Chorn, C., & Georgopoulos, A. (2015). White Noise Test : detecting autocorrelation and nonstationarities in long time series after ARIMA modeling. Proceedings Of The Python In Science Conferences. https://doi.org/10.25080/majora-7b98e3ed-00f

Marquardt, D.W. (1970) Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, 12, 591-612. https://doi.org/10.2307/1267205

McCracken, M.W., Ng, S., 2020. Fred-qd: A Quarterly Database for Macroeconomic Research, 005. Federal Reserve Bank of St. Louis Working Paper, pp. 1–52.

Meng, M., & Niu, D. (2011). Modeling CO2 emissions from fossil fuel combustion using the logistic equation. Energy, 36(5), 3355-3359. Online https://doi.org/10.1016/j.energy.2011.03.032, consulted on 7 July 2024

Moffat, I.U. and Akpan, E.A. (2019) White Noise Analysis: A Measure of Time Series Model Adequa- cy. Applied Mathematics, 10, 989-1003. https://doi.org/10.4236/am.2019.1011069

Muth, J.F. (1960). Optimal properties of exponentially weighted forecasts. Journal of the American Statistical Association, **55** (290), 299-306.

National Development and Reform Commission (NDRC) People's Republic of China. (2030). ACTION PLAN FOR CARBON DIOXIDE PEAKING BEFORE 2030 Online https://en.ndrc.gov.cn/policies/202110/t20211027_1301020.html, consulted on 10 July 2024

Nguyen, D. K., Huynh, T. L. D., & Nasir, M. A. (2021). Carbon emissions determinants and forecasting : Evidence from G6 countries. Journal Of Environmental Management, 285, 111988. https://doi.org/10.1016/j.jenvman.2021.111988

Ning, L., Pei, L., & Li, F. (2021). Forecast of China's Carbon Emissions Based on ARIMA Method. Discrete Dynamics In Nature And Society, 2021, 1-12. https://doi.org/10.1155/2021/1441942

Pérez-Suárez, R., & López-Menéndez, A. J. (2015). Growing green ? Forecasting CO2 emissions with Environmental Kuznets Curves and Logistic Growth models. Environmental Science & Policy, 54, 428-437. Online https://doi.org/10.1016/j.envsci.2015.07.015, consulted on 22 June 2024

Quéré, C. L., Andrew, R. M., Friedlingstein, P., Sitch, S., Hauck, J., Pongratz, J., Pickers, P. A., Korsbakken, J. I., Peters, G. P., Canadell, J. G., Arneth, A., Arora, V. K., Barbero, L., Bastos, A., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Doney, S. C.,. . . Zheng, B. (2018). Global Carbon Budget 2018. Earth System Science Data, 10(4), 2141-2194. https://doi.org/10.5194/essd-10-2141-2018

Overview of Greenhouse Gases | US EPA. (2024, 11 avril). US EPA. Online https://www.epa.gov/ghgemissions/overview-greenhouse-gases, consulted on 28 April 2024

Rahman, A., & Hasan, M. M. (2017). Modeling and Forecasting of Carbon Dioxide Emissions in Bangladesh Using Autoregressive Integrated Moving Average (ARIMA) Models. Open Journal Of Statistics, 07(04), 560-566. https://doi.org/10.4236/ojs.2017.74038

Rode, A., Carleton, T., Delgado, M., Greenstone, M., Houser, T., Hsiang, S., Hultgren, A., Jina, A., Kopp, R. E., McCusker, K. E., Nath, I., Rising, J., & Yuan, J. (2021). Estimating a social cost of carbon for global energy consumption. Nature, 598(7880), 308-314. https://doi.org/10.1038/s41586-021-03883-8

Saunois, M., Martinez, A., Poulter, B., Zhang, Z., Raymond, P., Regnier, P., Canadell, J. G., Jackson, R. B., Patra, P. K., Bousquet, P., Ciais, P., Dlugokencky, E. J., Lan, X., Allen, G. H., Bastviken, D., Beerling, D. J., Belikov, D. A., Blake, D. R., Castaldi, S.,. . . Zhuang, Q. (2024). Global Methane Budget 2000–2020. Earth Syst. Sci. Data Discuss. https://doi.org/10.5194/essd-2024-115

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6(2), 461–464. http://www.jstor.org/stable/2958889

Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 8(2), 39-42.

Singh, S., & Yadav, A. (2021). Interconnecting the environment with economic development of a nation. Elsevier eBooks(p. 35-60). https://doi.org/10.1016/b978-0-12-822188-4.00003-8

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58 (1), 267–288.

Tudor, C. (2016). Predicting the Evolution of CO2 Emissions in Bahrain with Automated Forecasting Methods. Sustainability, 8(9), 923. https://doi.org/10.3390/su8090923

United Nations. (s. d.). Net Zero Coalition | United Nations. Online https://www.un.org/en/climatechange/net-zero-coalition#:~:text=zero%20by%202050.-,2030,cent%20compared%20to%202010%20levels, consulted on 23 June 2024

World Development Indicators | DataBank. (s. d.). https://databank.worldbank.org/source/world-development-indicators

Yoro, K. O., & Daramola, M. O. (2020). CO2 emission sources, greenhouse gases, and the global warming effect. Dans Elsevier eBooks (p. 3-28). https://doi.org/10.1016/b978-0-12-819657-1.00001-3

7. Appendixes

Appendix 1.
R-squared between predictors and CO2 emissions in simple linear regression. All predictors are differentiated once, as the CO2 emissions of both countries.

Variables used for both countries analyses:

Variable	**R-squared China**	**R-squared India**
Agricultural land (sq. km)	0.019	0.005
Agriculture, forestry, and fishing, value added (constant LCU)	7.4e-05	0.034
Capture fisheries production (metric tons)	0.0009	0.005
Cereal production (metric tons)	0.0002	0.0001
Changes in inventories (current LCU)	0.149	0.074
Crop production index (2014-2016 = 100)	0.0186	0.003
DEC alternative conversion factor (LCU per US$)	0.002	0.0003
Exports of goods and services (current LCU)	0.037	0.016
External balance on goods and services (current LCU)	0.0002	0.203
Fertilizer consumption (kilograms per hectare of arable land)	0.066	0.018
Final consumption expenditure (current LCU)	0.118	0.112
GDP (constant LCU)	0.427	0.087
GNI (current LCU)	0.427	0.088
Gross capital formation (current LCU)	0.488	0.047
Gross domestic savings (current LCU)	0.369	0.085
Gross fixed capital formation (current LCU)	0.436	0.098
Gross national expenditure (current LCU)	0.466	0.122
Households and NPISHs Final consumption expenditure (current LCU)	0.0327	0.086
Imports of goods and services (current LCU)	0.109	0.045
Industry (including construction) - value added (constant LCU)	0.503	0.202
Inflation, GDP deflator (annual %)	0.086	0.012
Land under cereal production (hectares)	0.005	0.016
Merchandise exports (current US$)	0.049	0.046
Merchandise imports (current US$)	0.116	0.042
Official exchange rate (LCU per US$, period average)	0.002	2.3e-05
Permanent cropland (% of land area)	0.021	0.032
Renewable internal freshwater resources per capita (cubic meters)	0.031	0.037
Rural population (% of total population)	0.012	0.012
Services, value added (current LCU)	0.174	0.271
Total fisheries production (metric tons)	0.013	0.001
Trade (% of GDP)	0.004	0.012

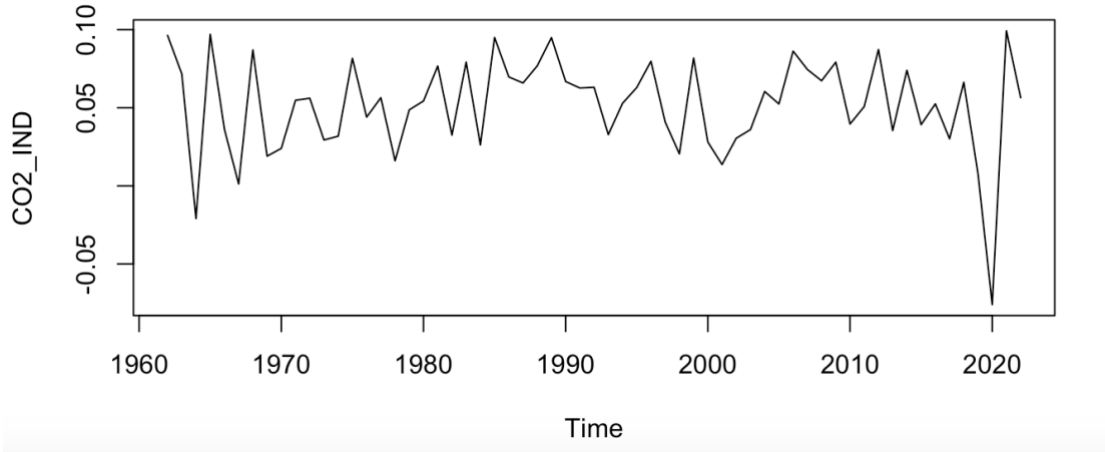
Aquaculture production (metric tons)	0.070	0.015
Arable land (hectares)	0.004	0.002

Variables used for India analyses only:

Variable	**R-squared India**
Exports as a capacity to import (constant LCU)	0.011
Inflation, consumer prices (annual %)	0.054

Appendix 2.

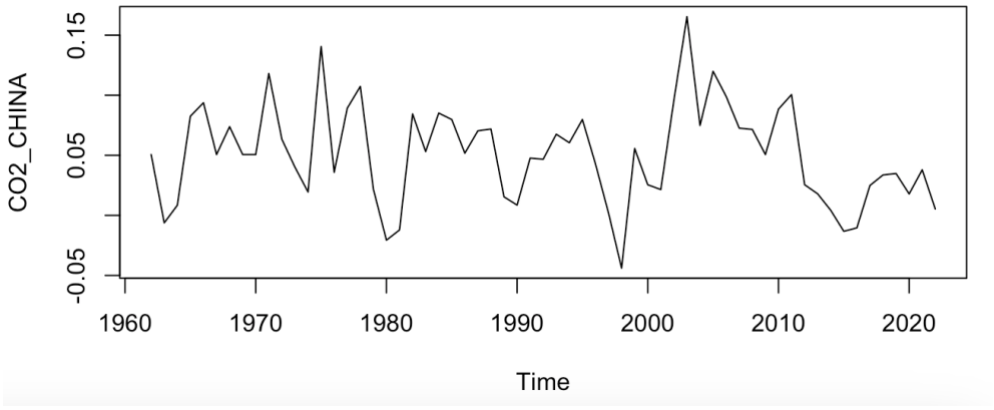
Plotting of India's CO2 emissions (differences of logarithms). Potential breaking point at n=58:




```
> sctest(CO2_IND ~ time1962, type="Chow", point= 58)

        Chow test

data:  CO2_IND ~ time1962
F = 6.3644, p-value = 0.0032
```

Appendix 3.

Plotting of China's CO2 emissions with outliers replaced(differences of logarithms). Potential breaking pointat n=14, n=35, n=45:




```
> sctest(CO2_CHINA ~ time1962, type="Chow", point= 14)

        Chow test

data:  CO2_CHINA ~ time1962
F = 1.1527, p-value = 0.323
```

```
> sctest(CO2_CHINA ~ time1962, type="Chow", point= 35)

        Chow test

data:  CO2_CHINA ~ time1962
F = 0.70247, p-value = 0.4996
```

```
> sctest(CO2_CHINA ~ time1962, type="Chow", point= 45)

        Chow test

data:  CO2_CHINA ~ time1962
F = 3.0543, p-value = 0.05494
```

Appendix 4.

ARIMA forecasting quality for China's CO2 emissions:

```
> accuracy(F_AR_CHN)
                       ME       RMSE        MAE       MPE     MAPE      MASE       ACF1
Training set -1.582845e-06 0.03787142 0.03045647 -47.08303 157.0682 0.8710011 0.01422581
```

Appendix 5.

ARIMA forecasting quality for India's CO2 emissions:

```
> accuracy(F_AR_IND)
                       ME       RMSE        MAE       MPE     MAPE      MASE       ACF1
Training set 3.15958e-05 0.02450047 0.02019076 -63.16059 100.3212 0.6313234 0.02331323
```

Appendix 6.

Outputs of LASSO selected variables after running the code 10 times (China):

```
> lambda_optimal_CHN
[1] 0.007866066
> # Fit of final LASSO model using the optimal lambda
> final_lasso_CHN <- glmnet(x_CHN, y_CHN, alpha = 1, lambda = lambda_optimal_CHN)
> coef(final_lasso_CHN)
24 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  0.011465536
b           -0.226366832
c            .
d            .
e            .
f            .
g            .
h            .
i            0.186172772
j            .
k            .
l            .
m            .
n            0.310428217
o            .
p            0.001976933
q            .
r            .
s            .
t            .
u            .
v            .
w            .
x            .
```

```
> lambda_optimal_CHN
[1] 0.005950391
> # Fit of final LASSO model using the optimal lambda
> final_lasso_CHN <- glmnet(x_CHN, y_CHN, alpha = 1, lambda = lambda_optimal_CHN)
> coef(final_lasso_CHN)
24 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  0.011844113
b           -0.322173755
c            .
d            .
e            .
f            .
g            .
h            .
i            0.205017545
j            .
k            .
l            0.002807748
m            .
n            0.312498961
o            .
p            0.016804932
q            .
r            .
s            .
t            .
u            .
v            .
w            .
x            .
```

```
> lambda_optimal_CHN
[1] 0.007167266
> # Fit of final LASSO model using the optimal lambda
> final_lasso_CHN <- glmnet(x_CHN, y_CHN, alpha = 1, lambda = lambda_optimal_CHN)
> coef(final_lasso_CHN)
24 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  0.011642326
b           -0.261275116
c            .
d            .
e            .
f            .
g            .
h            .
i            0.193303110
j            .
k            .
l            .
m            .
n            0.311401628
o            .
p            0.007361681
q            .
r            .
s            .
t            .
u            .
v            .
w            .
x            .
```

Appendix 7.

Linear regression between fossil CO2 emissions and predictors of three LASSO sets (China):

Model 1:

```
Call:
lm(formula = l_CO2_CHN ~ l_agri_CHN + l_industry_CHN + l_cap_form_CHN)

Residual standard error: 0.0467 on 55 degrees of freedom
Multiple R-squared:  0.6021,    Adjusted R-squared:  0.5804
F-statistic: 27.75 on 3 and 55 DF,  p-value: 4.631e-11
```

Model 2:

```
Call:
lm(formula = l_CO2_CHN ~ l_agri_CHN + l_industry_CHN + l_cap_form_CHN +
    l_merch_exp_CHN)

Residual standard error: 0.04655 on 54 degrees of freedom
Multiple R-squared:  0.612,     Adjusted R-squared:  0.5832
F-statistic: 21.29 on 4 and 54 DF,  p-value: 1.388e-10
```

Model 3:

```
Call:
lm(formula = l_CO2_CHN ~ l_agri_CHN + l_industry_CHN + l_cap_form_CHN +
    l_merch_exp_CHN + l_gne_CHN)

Residual standard error: 0.04693 on 53 degrees of freedom
Multiple R-squared:  0.6129,    Adjusted R-squared:  0.5764
F-statistic: 16.78 on 5 and 53 DF,  p-value: 6.534e-10
```

Appendix 8.

VIF values for best LASSO model (China):

```
> vif(lm(l_CO2_CHN ~ l_agri_CHN+l_industry_CHN+ l_cap_form_CHN +l_merch_exp_CHN))
     l_agri_CHN  l_industry_CHN  l_cap_form_CHN l_merch_exp_CHN
       1.148714        3.300642        3.442234        1.048010
```

Appendix 9.

Outputs of LASSO selected variables after running the code 10 times (India):

```
> lambda_optimal_IND
[1] 0.00572002
>
> # Fit of final LASSO model using the optimal lambda
> final_lasso_IND <- glmnet(x_IND, y_IND, alpha = 1, lambda = lambda_optimal_IND)
> coef(final_lasso_IND)
25 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  4.116187e-02
b           -1.917394e-02
c            .
d            .
e            .
f            .
g            .
h           -3.070248e-13
i            .
j            .
k            .
l            .
m            .
n            1.591158e-02
o            .
p            .
q            .
r            .
s            1.833361e-01
t            .
u            .
v            .
w            .
x            .
y            .
```

```
> lambda_optimal_IND
[1] 0.007561529
>
> # Fit of final LASSO model using the optimal lambda
> final_lasso_IND <- glmnet(x_IND, y_IND, alpha = 1, lambda = lambda_optimal_IND)
> coef(final_lasso_IND)
25 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  4.314132e-02
b            .
c            .
d            .
e            .
f            .
g            .
h           -2.319661e-13
i            .
j            .
k            .
l            .
m            .
n            3.637140e-03
o            .
p            .
q            .
r            .
s            1.419547e-01
t            .
u            .
v            .
w            .
x            .
y            .
```

```
> lambda_optimal_IND
[1] 0.002256094
>
> # Fit of final LASSO model using the optimal lambda
> final_lasso_IND <- glmnet(x_IND, y_IND, alpha = 1, lambda = lambda_optimal_IND)
> coef(final_lasso_IND)
25 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  3.207081e-02
b           -1.657060e-01
c            .
d            .
e            .
f            2.510158e-02
g            2.781441e-02
h           -3.748500e-13
i           -1.054328e-03
j            2.072923e-01
k            .
l            .
m            .
n            4.199608e-02
o            .
p            .
q            .
r            .
s            2.094615e-01
t           -5.105264e-04
u            .
v            .
w            .
x            .
y            .
```

Appendix 10.

Linear regression between fossil CO2 emissions and predictors of three LASSO sets (India):

Model 1:

```
Call:
lm(formula = l_CO2_IND ~ l_agri_IND + l_ext_balance_ADJ_IND +
    l_industry_IND + l_dom_sav_IND)

Residual standard error: 0.02539 on 54 degrees of freedom
Multiple R-squared:  0.3848,    Adjusted R-squared:  0.3393
F-statistic: 8.446 on 4 and 54 DF,  p-value: 2.286e-05
```

Model 2:

```
Call:
lm(formula = l_CO2_IND ~ l_agri_IND + l_final_cons_exp_ADJ_IND +
    l_industry_IND + l_dom_sav_IND + l_dec_chg_IND)

Residual standard error: 0.02495 on 53 degrees of freedom
Multiple R-squared:  0.4166,    Adjusted R-squared:  0.3616
F-statistic:  7.57 on 5 and 53 DF,  p-value: 2e-05
```

Model 3:

```
Call:
lm(formula = l_CO2_IND ~ l_ext_balance_ADJ_IND + l_industry_IND +
    l_dom_sav_IND)

Residual standard error: 0.02626 on 55 degrees of freedom
Multiple R-squared:  0.3297,    Adjusted R-squared:  0.2932
F-statistic: 9.018 on 3 and 55 DF,  p-value: 5.942e-05
```

Appendix 11.

VIF values for best LASSO model (India):

```
vif(lm(l_CO2_IND ~ l_agri_IND+ l_final_cons_exp_ADJ_IND+l_industry_IND+l_dom_sav_IND+l_dec_chg_IND))
   l_agri_IND l_final_cons_exp_ADJ_IND          l_industry_IND           l_dom_sav_IND           l_dec_chg_IND
     1.522544                2.247566                1.699459                1.169677                1.081318
```

Appendix 12.

Cointegration relationships for Granger set, China:

```
Browse[1]> summary(vecm.test_CHN0)

######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , with
linear trend in cointegration

Eigenvalues (lambda):
[1] 2.935702e-01 1.792716e-01 5.551115e-17

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 1 | 11.66 10.49 12.25 16.26
r = 0  | 20.50 16.85 18.96 23.65
```

Appendix 13.

Cointegration relationships for LASSO set, China.

```
Browse[1]> summary(vecm.test_CHN)

######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , with
linear trend in cointegration

Eigenvalues (lambda):
[1] 6.211099e-01 4.471320e-01 3.486892e-01 3.103740e-01
[5] 1.448156e-01 2.229411e-18

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 4 |  9.23 10.49 12.25 16.26
r <= 3 | 21.92 16.85 18.96 23.65
r <= 2 | 25.30 23.11 25.54 30.34
r <= 1 | 34.97 29.12 31.46 36.65
r = 0  | 57.26 34.75 37.52 42.36
```

Appendix 14.

VECM equations India:

```
Browse[1]> vecm_CHN <- VECM(dataset_CHN, lag = 1, r = 1, estim = "ML")
Browse[1]> summary(vecm_CHN) #vecm model
#############
###Model VECM
#############
Full sample size: 61    End sample size: 59
Number of variables: 5  Number of estimated slope parameters 35
AIC -1813.887   BIC -1732.863   SSR 1.420335
Cointegrating vector (estimated by ML):
   CO2_CHINA      AGRI INDUSTRY  CAP_FORM  MERCH_EXP
r1         1 -3.551497 2.285619 -2.499729 -0.2519169


                   ECT                Intercept          CO2_CHINA -1        AGRI -1
Equation CO2_CHINA -0.0820(0.0626)     -0.0094(0.0092)     -0.1877(0.1513)     0.2215(0.2943)
Equation AGRI      0.0366(0.0281)      0.0044(0.0041)      -0.0450(0.0678)     -0.5039(0.1319)***
Equation INDUSTRY  -0.0065(0.0705)     -0.0029(0.0103)     -0.1519(0.1703)     0.1610(0.3312)
Equation CAP_FORM  0.3559(0.1148)**    0.0389(0.0168)*     -0.4079(0.2775)     1.2370(0.5395)*
Equation MERCH_EXP 0.1297(0.1786)      0.0157(0.0262)      -0.0156(0.4318)     0.4877(0.8397)
                   INDUSTRY -1         CAP_FORM -1         MERCH_EXP -1
Equation CO2_CHINA -0.0779(0.1969)     -0.0469(0.1254)     -0.0009(0.0485)
Equation AGRI      -0.1186(0.0882)     0.0970(0.0562).     -0.0029(0.0217)
Equation INDUSTRY  -0.4177(0.2216).    0.0442(0.1411)      0.0806(0.0546)
Equation CAP_FORM  -0.1765(0.3610)     0.1949(0.2298)      0.1982(0.0889)*
Equation MERCH_EXP 0.3682(0.5618)      -0.1637(0.3577)     -0.3916(0.1384)**
```

Appendix 15.

Cointegration relationships for LASSO set, India:

```
####################
# Johansen-Procedure #
####################

Test type: maximal eigenvalue statistic (lambda max) , with linear trend in cointegration

Eigenvalues (lambda):
[1] 0.6509286 0.5249415 0.5110624 0.4011289 0.2671258 0.2436025 0.0000000

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 5 | 16.47 10.49 12.25 16.26
r <= 4 | 18.34 16.85 18.96 23.65
r <= 3 | 30.25 23.11 25.54 30.34
r <= 2 | 42.22 29.12 31.46 36.65
r <= 1 | 43.91 34.75 37.52 42.36
r = 0  | 62.10 40.91 43.97 49.51
```

Appendix 16.

VECM equations India:

```
Browse[1]> vecm_IND <- VECM(dataset_IND, lag = 1, r = 3, estim = "ML")
Browse[1]> summary(vecm_IND) #vecm model
#############
###Model VECM
#############
Full sample size: 61    End sample size: 59
Number of variables: 6  Number of estimated slope parameters 60
AIC -2353.026   BIC -2209.676   SSR 0.4756851
Cointegrating vector (estimated by ML):
       CO2_INDIA          AGRI    FINAL_CONS    INDUSTRY DOM_SAVINGS     DEC_CHG
r1  1.000000e+00  0.000000e+00 -2.775558e-17 -0.00808082 -0.09445112 -0.41093975
r2 -6.245005e-17  1.000000e+00  0.000000e+00 -0.30915875 -0.09509255  0.07419298
r3 -5.551115e-17 -1.110223e-16  1.000000e+00 -0.02772121  0.60547855 -0.27548688


                     ECT1                ECT2                ECT3                Intercept
Equation CO2_INDIA   -0.5843(0.1822)**   0.1562(0.1585)      0.1838(0.1287)      0.0100(0.0103)
Equation AGRI        0.1049(0.2790)      -1.5585(0.2427)***  -0.3057(0.1970)     0.0387(0.0157)*
Equation FINAL_CONS  -0.2634(0.1575)     -0.3593(0.1369)*    -0.0939(0.1112)     0.0211(0.0089)*
Equation INDUSTRY    -0.6448(0.1973)**   -0.0736(0.1716)     -0.2958(0.1393)*    0.0374(0.0111)**
Equation DOM_SAVINGS 0.6500(0.3788).     0.1326(0.3295)      -1.6456(0.2675)***  0.0471(0.0213)*
Equation DEC_CHG     1.7060(0.3419)***   0.5227(0.2974).     0.0653(0.2414)      -0.0738(0.0192)***
                     CO2_INDIA -1        AGRI -1             FINAL_CONS -1       INDUSTRY -1
Equation CO2_INDIA   -0.2519(0.1544)     -0.0443(0.0954)     -0.1230(0.2083)     -0.0174(0.1482)
Equation AGRI        -0.2634(0.2364)     0.0710(0.1461)      0.1740(0.3190)      -0.1726(0.2270)
Equation FINAL_CONS  0.1238(0.1334)      0.1685(0.0825)*     -0.5464(0.1800)**   0.0537(0.1281)
Equation INDUSTRY    0.4792(0.1672)**    -0.0188(0.1033)     0.3069(0.2255)      -0.3057(0.1605).
Equation DOM_SAVINGS -0.1278(0.3210)     -0.1712(0.1984)     1.6509(0.4331)***   0.3282(0.3082)
Equation DEC_CHG     -0.7450(0.2897)*    -0.2205(0.1791)     0.2709(0.3908)      -0.1614(0.2782)
                     DOM_SAVINGS -1      DEC_CHG -1
Equation CO2_INDIA   -0.1460(0.0648)*    -0.0587(0.0682)
Equation AGRI        0.0893(0.0992)      0.0558(0.1044)
Equation FINAL_CONS  -0.0253(0.0560)     -0.0378(0.0589)
Equation INDUSTRY    -0.0245(0.0701)     -0.2020(0.0738)**
Equation DOM_SAVINGS 0.1284(0.1347)      -0.0622(0.1417)
Equation DEC_CHG     0.1118(0.1216)      -0.1033(0.1279)
```

Appendix 17.

Forecasting accuracy of VECM model, China:

```
Browse[1]> print(paste("MAE: ", mae_CHN))
[1] "MAE:  0.0566050754002813"
Browse[1]> print(paste("RMSE: ", rmse_CHN))
[1] "RMSE:  0.0680713450042878"
Browse[1]> print(paste("MAPE: ", mape_CHN, "%"))
[1] "MAPE:  104.193602253335 %"
```

Appendix 18.

Forecasting accuracy of VECM model, India:

```
Browse[1]> print(paste("MAE: ", mae_IND))
[1] "MAE:  0.0540674906676317"
Browse[1]> print(paste("RMSE: ", rmse_IND))
[1] "RMSE:  0.0617181100609306"
Browse[1]> print(paste("MAPE: ", mape_IND, "%"))
[1] "MAPE:  115.665612032842 %"
```