Impact of PM2.5 Pollution on China's Economy: A Regression and Time Series Analysis Carried Out On 11 Chinese Provinces/Municipalities.

Submitted by: Sana Rabia Khan

MSc Student in Data Science and Statistics

CONTENTS

1. Introduction

Since environmental pollution threatens both economic growth and human well-being, it is a crucial issue in the process of sustainable economic development. The multiple harmful consequences on health, resource depletion, and the natural disasters linked to climate change are all caused by pollution (Azam, 2016). The great majority of the population on earth is exposed to PM2.5 pollution levels that are higher than the 10 $\mu g/m^3$ annual mean (25 $\mu g/m^3$ 24-hour mean) Air Quality Guidelines (AQG) set by the World Health Organization (WHO) (Health Organization, 2016), (Liao et al., 2020). Numerous studies have linked exposure to high levels of PM2.5 with greater rates of outpatient visits and hospital admissions for respiratory, cardiovascular, and cerebrovascular disorders. Long-term or acute exposure to a serious air pollutant both have the potential to increase mortality. These health issues can have a significant negative impact on the economy by raising health care expenditure, causing more missed workdays, and reducing the supply of labour (Xie et al., 2016). China's economy has been growing quickly ever since "reform and opening up" began in earnest. But simultaneously, China's environmental quality has been gradually declining (Hao et al., 2018). Air pollution has caused economic losses in China and around the world in addition to causing health problems. Over the past few decades, China has experienced major air quality problems as a result of its fast industrialisation and urbanisation. A significant problem affecting China's environmental quality, public health, and economic sustainable development is the persistent and heavy haze weather, represented by PM2.5, which happens more frequently, on an unprecedented enormous scale, has become a major issue (Liao et al., 2020). With an annual average fine particulate matter (PM2.5) concentration of 93 micrograms per cubic metre ($\mu g/m^3$) in 2014, the Beijing-Tianjin-Hebei (Jing-Jin-Ji) region experienced particularly severe air pollution, far exceeding both China's national standard and the standard recommended by the World Health Organization (WHO). The Chinese Academy for Environmental Planning (CAEP) calculated in 2015 that the cities' respective PM2.5 emissions were significantly greater than their ability to absorb the environment by 80% (China: Fighting Air Pollution and Climate Change through Clean Energy Financing, 2020). It is projected that China would face a GDP loss of 2.0% under the WoPol scenario and health costs related to PM2.5 pollution of 25.2 billion USD in 2030. In comparison, the WPol scenario's control strategy projected a gain of 1.17% of China's GDP from reduced PM2.5 pollution vs a control investment of 101.8 billion USD (0.79% of GDP). Tianjin (3.08%), Shanghai (2.98%), Henan (2.32%), Beijing (2.75%), and Hebei (2.60%) have the biggest GDP losses at the provincial level in the WoPol scenario for 2030, whereas Henan, Sichuan, Shandong, Hebei, and Jiangsu have the highest increases in healthcare expenditures. In two-thirds of the provinces, reducing PM2.5 pollution could result in beneficial outcomes. Due to increased PM2.5 pollution and densely populated areas, Tianjin, Shanghai, Beijing, Henan, Jiangsu, and Hebei gain the most from PM2.5 pollution reduction (Xie et al., 2016).

With air pollution being a major problem for the economy, China has made significant efforts to avoid and manage air pollution, and recent years have seen an improvement in the country's air quality. Since 2013, China has started enforcing strict clean air policies in an effort to improve air quality (Xiao et al., 2020). Air pollution management policies have undergone significant changes, particularly a shift from weak to robust execution. Large-scale national air pollution control policies, such as the Air Pollution Prevention and Control Action Plan, have recently been launched and put into action (Jin, Andersson and Zhang, 2016). Some of the other policies were the Innovative Financing for Air Pollution Control in Jing-Jin-Ji Program (China - Innovative Financing for Air Pollution Control in Jing-Jin-Ji Project, 2016) and the introduction of new instruments to the 11th Five-Year Plan to reduce emissions etc (Jin, Andersson and Zhang, 2016). There are several studies conducted on the relationship and the impact of air pollution on public health, but there are very few researches made on the economic effects of China’s province level. This dissertation attempts to cover this gap as an analysis is carried out to evaluate the impact of PM2.5 pollution on the economy of 11 Chinese provinces. Although, there are numerous ways that can be used to assess this output, regression and time series models have been utilised in this paper to achieve the required results. Regression analysis was chosen as it offers incredible flexibility and may be applied in a wide range of situations. Multiple independent variables can be modelled using regression analysis, as can continuous and categorical variables, polynomial terms can be used to model curvature, and interaction terms can be evaluated. It can also unravel incredibly complex issues when the variables are intertwined (Frost, 2022). As the analysis performed in this report includes independent and dependent variables that is continuous in nature, this method was thought best applicable. On the other hand, a “time series analysis” is a particular method of examining a set of data points gathered over a period of time. The analysis provides more details and demonstrates how factors change over time. Classification, curve fitting, descriptive analysis, exploratory analysis, and forecasting are a few of the models used in time series analysis (Time Series Analysis: Definition, Types, Techniques, and When It's Used, 2022). We use this method of analysis to observe the trend and seasonality of the data over the years and forecast the PM2.5 pollution and the GDP of China’s provinces.

The main objective of this dissertation is to build a model that signifies the relationship between particulate matter that is less than 2.5 micrometers in diameter (PM2.5) and the economy of China's provinces. We have considered the values of gross domestic product (GDP) as it is the most common way to measure the growth of an economy. This is carried out to observe the impact PM2.5 has on China's economy. Another objective of building a model would be to examine the trend of the data and perform forecasting or predict the PM2.5 level and the growth of GDP in China. Airborne particulate matter (PM) is a composite of several chemical species rather than a single pollutant. It consists of a complex mixture of solids and aerosols, including dry solid particles, liquid-coated solid cores, and small liquid droplets. Particles can contain inorganic ions, metallic compounds, elemental carbon, organic compounds, and chemicals from the earth's crust. They can vary greatly in size, shape, and chemical composition (Inhalable Particulate Matter and Health (PM2.5 and PM10) | California Air Resources Board, 2022). When levels of fine particulate matter (PM2.5) in the air are excessive, it is a risk for people's health which in turn is a risk for the sustainability of the economy. When the PM2.5 levels are high it causes a decrease in visibility and makes the air appear hazy. Both indoor and outdoor sources can produce PM2.5. The exhaust from cars, trucks, buses, and off-road vehicles, as well as emissions from other operations that involve burning of fuels are some of the examples of outdoor sources. This also includes natural sources such as forest and grass fires (Fine Particles (PM 2.5) Questions and Answers, 2022). With the increase in urbanisation and industrialization which involves an increase in the fuel emissions, the level of PM2.5 also increases. This provides a threat to the growth of the economy. The government and policy makers must comprehend the significance of the relationship between the two. In order to take the appropriate actions regarding the issue, they need to understand whether the GDP values depend on the level of PM2.5 in the environment. An observation of the trend of the two variables is also equally important. With the help of regression and time series models, this result can be achieved by the relevant authorities. The regression model may produce the estimates of the variables under the assumption of the probability distribution using the quarterly GDP data gathered over a period of time along with the PM2.5 daily data, which is afterwards transformed to quarterly data. The time series model, however, can support forecasting.

This dissertation is organised as follows: An overview of the many methods and models that have been used in the past to calculate the impact of fine particulate matter on the economy is given in Chapter 2. Some of the examples also cover another aspect to this thesis topic. The third chapter goes on to describe the study areas and data used in this thesis, as well as the different types of methods and the model's basic framework. After going over the framework, each method's benefits and drawbacks are discussed, along with any relevant key assumptions. GAM and ARIMA models are developed using the framework from Chapter 3; the resulting outcomes are examined in Chapter 4. The major steps in constructing the model are then summarised, noteworthy discoveries are highlighted, and additional forecasting improvements as well as overall improvements to the models are suggested in the conclusion.

2. Literature Review

This chapter talks about the various common methods or approaches that have been used in analysing the relationship between GDP and pollutants and estimating them. The methods include Geographically weighted regression model (GWR), Integrated model of Energy, Environment and economy for sustainable Development/Computable General Equilibrium model (IMED/CGE), Simultaneous Equations Model (SEM), Empirical Analyse and Instrumental Variables approach and Spatial Durbin Model (SDM). All of the strategies covered in this chapter need progressively more modelling, data collection, and equipment. Each method is briefly described, and benefits and drawbacks or limitations for most of these approaches are covered. Applications from published studies are shown when they are crucial to comprehend. Some of these studies also examine how PM2.5 pollution affects GDP in a way that isn't covered in this thesis. This is done in order to view the topic from a wider angle.

2.1 Geographically Weighted regression model (GWR)

As per previous studies conducted to analyse the correlation between the economic growth (GDP) and the PM2.5 pollution, an assumption was generally made that the relationship does not differ with the spatial position (Yan et al., 2021) . But (Yan et al., 2021) considered the geographical aspect of it in his study. He used Geographically weighted regression model (GWR) to analyse the socioeconomic factors affecting the urban PM2.5 pollution in China. Hong Kong, Macao, Taiwan, Tibet, Qinghai Province, and other places with missing data not included in the study region. This model is a local variable coefficient model that Fotheringham suggested for detecting spatial nonstationarity (Fotheringham, Brunsdon and Charlton, 2010). Geographically weighted regression (GWR) is a geographical analysis technique that models the local associations between these predictors and an outcome of interest while taking into account non-stationary variables (such as climate, demographic factors, and physical environment features) (Geographically Weighted Regression | Columbia Public Health, 2022). This study used the PM2.5 yearly mean concentrations data and eight anthropogenic perspectives amongst which GDP was one of it. GDP here referred to the level of economic development. The main aim of using this model was to conduct cross-sectional analysis of spatial data, which can more thoroughly and objectively reflect the changes in the socioeconomic factors over time in each city. In addition to reflecting the overall state of the regression relations of socioeconomic factors, the output results of the GWR model also reflected the specific regression situation of various components in each city. The results were provided in two parts: Global Regression results for GDP's influence on PM2.5 was observed to be positive and the geographical variation analysis resulted in socio-economic factors including GDP's influence on PM2.5 to vary among cities. This result seems to differ as per the output produced in this study as the relationship between GDP and the PM2.5 was observed to be negative. Although the focus of our study is on how the PM2.5 pollutant influences economic growth (GDP), it is important to consider how the results may alter if the independent and dependent variables were reversed. We reviewed this analysis after taking into account this change and looking at the geographical component of it.

Similar to other analytical techniques, GWR has a number of drawbacks, such as the multicollinearity of local coefficients, the inability to do numerous hypothesis tests, and the inability to break down the global estimates into local estimates. GWR is still thought of as a valuable method for investigating spatial non-stationarity and interpolation despite concerns (Chen, Deng, Yang and Matthews, 2012).

2.2 Integrated Model of Energy, Environment and economy for sustainable Development/Computable General Equilibrium model (IMED/CGE)

Another aspect of analysing the effect of PM2.5 on the economic growth is the health impact. As it is widely known that PM2.5 causes health problems which also leads to causing financial losses in China and around the world. This is usually referred to as health-related economic losses (Wang, Zhang, Niu and Liu, 2020). Previously, there have been many researches done on the health and economic impacts of PM2.5 pollution. One such study made by (Xie et al., 2019) uses an integrated approach to assess the effects of ambient air pollution, including the PM2.5 pollutant on China's economy and health. There were four scenarios that were implemented in this study —the reference, woPol, wPol, and wPol2. The reference scenario disregards the negative effects of air pollution on health, the expense of healthcare, preventable deaths, and lost workdays. Despite the fact that this scenario was thought of as ideal, its role was compared to that of the other scenarios. An assessment was then made on the negative impacts of pollution and the advantages of pollution control. IMED/CGE model was developed to provide the GAINS model information such as population growth, GDP growth rate, energy consumption and air pollutant emission by province and sector. This helped in quantifying the economic impacts of health damage. This model uses data of 30 provinces in China for the year 2002. The economic findings of the study stated that PM2.5 causes a 0.6% GDP loss in the wPol scenario and a 2.3% GDP loss in the woPol scenario and the health costs will total upto 210 billion CNY in 2030 (Xie et al., 2019). IMED is a system of databases and models that deals with the interconnected systems of energy, environment, and economy. Its goal is to quantitatively and systematically analyse economic, energy, environmental, and climate policies at the local, provincial, national, and international levels in order to provide decision-makers with the most up-to-date scientific information possible (IMED Overview, 2022). On the other hand, large-scale numerical models known as CGEs integrate economic theory with actual economic data to compute the effects of economic shocks or policy changes. The interdependencies between various economic sectors, agents, and markets are taken into consideration by CGE models. CGE analysis can therefore provide light on how policies affect the whole economy and occasionally show their side effects or unforeseen consequences. These models capture the entire economy and take into consideration interactions and knock-on effects across its many parts, as opposed to partial equilibrium models, which concentrate on a single sector of the economy (Computable General Equilibrium modelling: introduction, 2022). For instance, (Wang et al., 2016) in his article which assesses health and economic effects by PM2.5 pollution in Beijing uses CGE approach. He uses this model to assess the economic loss by taking into account two conducting variables, i.e.; the labour loss and excess medical costs. The findings indicate that Beijing's GDP loss as a result of resident health issues was estimated to be 1286.97 (95% CI: 488.58-1936.33) million RMB. Thus, it was clear that PM2.5 pollution has a negative impact on inhabitants' health and quality of life, as well as their ability to grow their economy. In yet another study that was previously conducted on the "Health and economic impacts of air pollution in China" (Hong-Wei, Toshihiko and Yue, 2006) makes use of CGE model to calculate the corresponding economic effects from a national perspective. 39 sectors and 32 commodities were involved in the construction of the model. As per the outcome of this approach, China's GDP loss was estimated at 0.38 per thousand (ranging from 0.16 per thousand to 0.51 per thousand).

Although CGE models' key advantage is its flexibility as it allows for the simulation of a wide range of policies and shocks (Computable General Equilibrium modelling: introduction, 2022), it has a few drawbacks as well. The use of CGE models is difficult, and the outcomes are heavily reliant on important economic characteristics that are still subject to uncertainty. These models are also costly and time-consuming (it takes months to years to build a CGE model). Because they don't adequately account for the actual behaviour of economic agents, CGE models are highly debatable. Indeed, the traditional economics that CGE models are based on has come under heavy scrutiny from academics across a variety of disciplines. But (Computable general equilibrium - Coastal Wiki, 2022) mentions that an individual is unable to make the best choices in order to accomplish its objectives using these models. He is only able to make decisions that are satisfying because its capacity, its habits and unconscious reflexes are a limitation for him. His values and conceptions of the end goal (which may even differ from the end objective determined by the enterprise); and his knowledge and the incomplete information he has access to could result in being a major drawback of CGE models if he unable to understand this (Computable general equilibrium - Coastal Wiki, 2022).

2.3 Simultaneous Equations Model (SEM)

A model in the form of a collection of simultaneous linear equations is known as a Simultaneous Equation Model (SEM). SEM models feature two or more equations, unlike models with a single equation that are introduced in introductory regression analysis (such as simple linear regression). Changes in the explanatory variable (X) in a single equation model generate changes in the response variable (Y); in a SEM model, additional Y variables are among the explanatory variables in each SEM equation. In other words, the system exhibits some sort of simultaneity or "back and forth" causation between the X and Y variables. The system is jointly driven by the equations in the system (Simultaneous Equations Model (SEM): Simple Definition, 2022). SEM is yet another model that is pretty commonly used to analyse the impact of air pollution including PM2.5 pollutant on the economy. Previously, not a lot of research has been made on this topic. (Hao et al., 2018) study on the impact of PM2.5 concentrations on per capita GDP is one of the relatively few studies on this subject. The economic impact of haze pollution in China was examined in this study using city-level panel data for the years 2013 to 2015 created from recently released urban PM2.5 concentrations. The estimations were carried out using the 3SLS approach, and a carefully built SEM that contained a pollution equation and a growth equation used to address the endogeneity that may have brought on by the bilateral causality between economic development and air pollution. Pooled data from all cities with information and panel data from the 73 cities that originally reported PM2.5 values were utilised to assure the stability and robustness of the estimation results. A number of time dummies and regional dummies were also added to the panel data analysis to compensate for the geographical and time fixed effects. The estimation findings showed that haze pollution did have a negative impact on China's economic growth (GDP) (Hao et al., 2018). This study's findings appeared to be consistent with the findings of this dissertation, where the negative impact by PM2.5 on the economy appeared to have also been noted.

There are a few presumptions that apply to these models. The endogenous and exogenous variables in simultaneous equation models are presumed to be directly measured and error-free. All variables that affect y but are left out of the equation and are considered to have expected values of zero are included in the disturbances. Furthermore, it is expected that disturbances are homoscedastic, nonautocorrelated, and uncorrelated with the exogenous variables. It is assumed that the random variables do not instantly affect one another (CHAPTER 2. SPECIFICATION OF SIMULTANEOUS EQUATION MODELS, 2022).

2.4 Empirical Analyse and Instrumental Variables Approach

A type of research known as empirical analysis is focused on locating concrete, verifiable evidence. Empirical analysis, which is guided by the scientific method, enables researchers to eliminate personal bias and instead use concrete, accurate, and repeatable real-world facts to derive conclusions. Empirical analysis' main idea is that the best approach to study reality and discover the truth is through direct observation (Empirical Analysis: Definition, Characteristics and Stages, 2021). One of the quantitative methods of an empirical analysis is correlational research. Finding relationships between two sets of variables is done through this research. In most cases, regression is utilised to forecast the results of such a method. Correlation can be either positive, negative, or neutral (Empirical Research: Definition, Methods, Types and Examples | QuestionPro, 2022). This correlational method was used by (Dong, Xu, Shen and He, 2021) in his study. On the basis of a sample from 2002 to 2017, this study empirically investigated the impact of air pollution on regional economic growth in Chinese provinces.

Fine particulate matter (PM2.5) concentration was used to measure air pollution, while the annual growth rate of GDP per capita was used to measure economic growth. For quantitative analysis, a panel data fixed-effects regression model was constructed using the instrumental variables estimation approach. In observational research, instrumental variables (IVs) are employed to account for confounding and measurement error. They make it possible to draw conclusions about causes from observational data (An Introduction to Instrumental Variables, n.d.). The research hypothesis in this study was confirmed as a consequence of the results, which clearly demonstrated how air pollution has a considerable detrimental impact on economic growth. It was calculated that a 1% rise in PM2.5 concentration causes the yearly growth rate of real GDP per capita to decrease by 0.05818 percentage points. According to the study, air pollution has a major negative impact on China's macroeconomic growth (Dong, Xu, Shen and He, 2021).

There are few benefits to an empirical approach. It greatly enhances the value of the research paper because it adds to the existing knowledge. This approach is also flexible and changes can be made as needed to the sample size, sampling style, data collection techniques, and analysis techniques. As they can be easily incorporated, fewer rules are followed. This approach is also time saving. However, this method comes with a few drawbacks as well. Empirical research takes a lot of time. Primary data analysis cannot be summarized into less than 3000 words, depending on the quantity of variables and the data analysis techniques employed. Results can be unforeseen. However, if the right steps are made in advance, this issue can be avoided. The display of data can be challenging because there is no standard format for the presentation of empirical data (Chetty, 2022).

2.5 Spatial Durbin Model (SDM)

In the field of spatial econometrics, the spatial Durbin model holds an interesting position. It is a model with cross-sectional dependence in the errors that has been reduced, and it can be used as the nesting equation in a more comprehensive model selection strategy (Mur and Angulo, 2006). This model was used by (Ding et al., 2019) in his study that was conducted to examine the existence of environmental Kuznets curve (EKC) of satellite observations of PM2.5 pollution in the Beijing-Tianjin-Hebei (BTH) region of China. In order to examine the association between economic growth and PM2.5 pollution, panel data was used from 13 cities in the BTH region from 1998 to 2016 along with satellite observations of PM2.5 pollution to conduct this study. In the investigation of EKC, the spatial Durbin Model (SDM) was preferred over the non-spatial model after confirming the spatial effect of PM2.5 pollution using spatial statistical analysis. In order to integrate spatially lagged independent variables and spatially autoregressive processes in the error term, the spatial Durbin error model (SDEM) was also created. The findings revealed a substantial inverted U-shaped pattern in the association between economic development and PM2.5 pollution.

The use of spatial Durbin model has some advantages to it. One of them being the benefit of producing unbiased coefficient estimates even when a spatial lag or a spatial error model is used to generate the actual data. Another advantage is that it does not place limitations on how much potential spatial spillover effects can be. These spillover effects, in contrast to previous spatial regression definitions, can be local or global and vary depending on the explanatory variable (Applied Spatial Econometrics: Raising the Bar, 2022).

3. Material and methods

3.1 Study Areas

As mentioned earlier in chapter 1, fine particulate matter (PM2.5) is one of the principal air contaminants in China due to the rapid industrialisation and increase in energy consumption (Lin, Zou, Yang and Li, 2018). With 1.4 billion people as of 2019, China has the largest population in the world. With a total land size of 9.6 million square kilometres, it ranks as the fourth-largest nation in the world. China was listed as the 11th dirtiest nation in the world in 2019. The PM2.5 pollution concentration was three times higher than what the World Health Organization (WHO) recommends (China Air Quality Index (AQI) and Air Pollution information | IQAir, 2022). In light of these facts, China was chosen as the area of study. This study considers China’s seven provinces and four municipalities at random to perform air pollution analysis. The seven provinces that were chosen at random were Hebei, Hubei, Jiangsu, Jilin, Shandong, and Zhejiang and the four municipalities were Beijing, Tianjin, Shanghai and Chongqing. A brief information is given below about these regions.

Beijing: The People's Republic of China's capital is Beijing which is a provincial-level shi (municipality). With the exception of two short stretches that border Tianjin municipality to the southeast, Beijing, the larger municipality is almost entirely encircled by Hebei province. Geologists refer to this concave arc that circles the Beijing plain from the northeast to the southwest as the "Bay of Beijing" since the Yan Mountains are located along the municipality's northeastern border and the Jundu Mountains take up its entire western region (Beijing - People, 2022). Beijing has constructed an economic structure that includes high-quality, precise, and advanced industries, attaining a combined GDP of RMB 3.6 trillion (US$537.3 billion), similar to the average level observed by industrialised countries, with a compound annual GDP growth rate of 6% between 2016 and 2020 (Briefing, 2022). But with the growth of industrialisation the air quality in Beijing declined. This caused an increased in the concentration of PM2.5 pollution (Xu and Zhang, 2020).

Tianjin: Tianjin is also a city and province-level shi (municipality) just like Beijing. It is situated at the northernmost point of the North China Plain, east of the province of Hebei. It is China's third-largest municipality after Shanghai and Beijing. Additionally, it serves as North China's principal port and most significant manufacturing hub (Tianjin | History, Map, Population, & Facts, 2022). According to (Tianjin Air Quality Index (AQI) and China Air Pollution | IQAir, 2022), Tianjin has a large proportion of road freight, and motor vehicle emissions—particularly those from diesel trucks, which make up less than 10% of vehicle ownership but account for 70% of all vehicle emissions—significantly contribute to nitrogen oxides in the atmosphere. Heavy-duty diesel vehicles make up over two thirds of the moving freight vehicles. Tianjin is also in the midst of an essential period of development and construction. Urban infrastructure development and other construction projects are dispersed widely and with great intensity (Tianjin Air Quality Index (AQI) and China Air Pollution | IQAir, 2022).

Shanghai: Shanghai is a significant industrial and economic hub for China as well as one of the largest seaports in the world. The city is located on the coast of the East China Sea between the mouth of the Yangtze River (Chang Jiang) to the north and the bay of Hangzhou to the south. The city itself, the neighbouring suburbs, and an agricultural hinterland are all included in the municipality's boundaries. Although Beijing is the capital of China, Shanghai is the largest city. It's also one of the most populated cities in China, and the municipality is the country's most populous urban area. Since the communists' triumph in 1949, it has developed into an economic powerhouse whose goods meet China's expanding domestic demand. Due to a unique combination of factors, Shanghai has long been the nation's premier industrial and manufacturing hub. The presence of a large, highly educated, and technologically innovative labour force, a solidly established and broadly based scientific research establishment that supports industry, a history of producer cooperation, excellent internal and external communication and supply facilities, and others are among them (Shanghai - Economy, 2022). Shanghai's air pollution is mostly brought on by the burning of coal, automobiles, industrial dust, atmospheric chemical reactions in metropolitan areas, and unfavourable meteorological conditions, all of which are associated with the city's rapid socioeconomic development (Shanghai Air Quality Index (AQI) and China Air Pollution | IQAir, 2022).

Chongqing: The city is situated at the confluence of the Yangtze and Jialing rivers, some 1,400 miles (2,250 km) from the sea. It serves as the main river port, a transportation hub, and the commercial and industrial heart of the upper Yangtze River (Chang Jiang) basin. It is the fourth municipality to be established after Beijing, Shanghai and Tianjin. Because of the accessibility of commodities like coal, iron, and other materials, industry grew quickly. The communist government invested significantly in industrial development after 1949. Chongqing was one of the largest and fastest-growing industrial hubs in southwest China by the late 20th century (Chongqing - Economy, 2022). There was increase demand for energy as these practices grew. This demand was met with the import of cheap coal from northern China which was of poor quality. This lead to a higher concentration of PM2.5 pollutant along with the other pollutants like sulphur dioxide (SO_2) (Chongqing Air Quality Index (AQI) and China Air Pollution | IQAir, 2022).

Hebei: Located on the Bo Hai (Gulf of Chihli) of the Yellow sea, Hebei is a province of Northern China. In addition to the provinces of Liaoning to the northeast, Shandong to the southeast, Henan to the south, and Shanxi to the west, it is bordered to the northwest by the Inner Mongolia Autonomous Region. Hebei is one of the most developed provinces in northern China in terms of both culture and economy (Hebei - Climate, 2022). As of 2019, it was the region that contributed the most to air pollution, accounting for 70% of PM 2.5 emissions. The highest yearly average PM 2.5 concentrations are found in this province. This is mainly due to a significant concentration of highly polluting industry, cars, and an extensive agricultural sector (China's Hebei Province Fights for Blue Skies with World Bank Support, 2019).

Jilin: China's Northeastern area includes the province of Jilin (formerly called Manchuria). It shares boundaries with Russia to the east, North Korea to the southeast, the Chinese provinces of Heilongjiang and Liaoning to the north, the Inner Mongolia Autonomous Region to the west, and Russia to the east. Changchun, located in the province's west central region, serves as the capital. Jilin is a major manufacturer of autos, chemicals, machine tools, power, and forest products and has a relatively high level of industrialization. The majority of the province's industry is centred in Changchun and Jilin, the two biggest cities (Jilin - People, 2022). This contributes to the rise in the emission of PM2.5.

Hubei: Hubei, which means “North of the lake”, is a province in central China that is a part of the Yangtze River's middle basin (Chang Jiang). Hubei is surrounded by the provinces of Shaanxi to the northwest, Henan to the north and northeast, Anhui to the east, and Jiangxi to the southeast. It also shares a border with the municipality of Chongqing to the west. Its capital is Wuhan, a composite name for the three former cities (now districts) of Hankou, Hanyang, and Wuchang. These three locations are halfway between Shanghai and Chongqing and are located at the confluence of the Han River and the Yangtze River, about 600 miles (965 km) from the sea. One of China's most significant industrial hubs at present are in Wuhan. Huangshi has expanded into a significant iron and steel hub. Together, the cities of Xianfan, Wuhan, and Shiyan in northwest Hubei have grown to be important national hubs for the automobile industry. In addition, a large shipyard is located in Wuhan's Wuchang district, and the province is home to numerous foreign-built factories that make chemical fertilisers (Hubei - Resources and power, 2022). Wuhan, one of China's most polluted cities, was ranked 146 internationally in the mid of 2019. The northern region is still experiencing severe pollution. "Fireworks and firecrackers have caused a large increase in pollutants," according to experts at the National Centre for Air Pollution Prevention and Control, and "highly polluting steel, coking, glass, refractory materials, chemical, pharmaceutical, and other heavy chemical industries have a large number of uninterruptible processes."(Wuhan Air Quality Index (AQI) and China Air Pollution | IQAir, 2022)

Guangdong: South China's Guangdong province is the most southern of the mainland's provinces and the area through which South China's trade is mainly routed. Among all provinces, Guangdong has one of the longest coasts, fronting the South China Sea to the southeast and south (including connections to the two special administrative regions of Hong Kong and Macau). In addition, it is surrounded by the provinces of Hunan, Jiangxi to the north and to the northeast by Fujian, the Zhuang Autonomous Region of Guangxi to the west. Guangzhou (Canton), located near the head of the Pearl (Zhu) River Delta, is the country's capital. Guangdong saw modern expansion during the first part of the 20th century as Guangzhou became a hub for commerce, industry, and transportation. Nevertheless, Guangdong received very little attention during the First Five-Year Plan due to the scarcity of its iron reserves (1953–57). However, the discovery of more mineral deposits lead to the growth of several more substantial businesses, such as shipbuilding and ship repair, metal and petrochemical processing, and machinery manufacturing. These industries are still heavily concentrated in Guangzhou (Guangdong - Settlement patterns, 2022). The rigidity of pollution emissions has increased along with the swift economic and social development as the megacity of Guangzhou has a large economy, a dense population, and a large source of overall pollution emissions. This makes it difficult to improve air quality. Guangzhou had bad air quality with a US AQI score of 149 around the end of 2020. This classification is based on the World Health Organization's (WHO) suggested levels. The amount of PM2.5 pollutant suspended in the air was 55 $\mu g/m^3$ (microns per cubic metre) (Guangzhou Air Quality Index (AQI) and China Air Pollution | IQAir, 2022).

Shandong: The northernmost coastal province of China, Shandong, which translates to "East of Mountains," is located across the Yellow Sea from the Korean peninsula. Only Henan has a larger population in China than the second-most populated province, Shandong. There are two distinct parts to the province. The first is an interior region bordered to the north and west by the province of Hebei, to the southwest by the province of Henan, and to the south by the provinces of Anhui and Jiangsu. The Shandong Peninsula is the second, stretching 200 miles (320 km) seaward from the Wei and Jiaolai river basins, with the Bo Hai (Gulf of Chihli) to the north and the Yellow Sea to the south. The peninsula makes up a significant portion of the province's 1,575 miles of coastline (2,535 km). Agricultural and manufacturing sectors of Shandong's economy are diverse. The main manufacturing hub of Qingdao is home to a large textile industry, a locomotive factory, as well as chemical, tire, and machine tool manufacturers. Large-scale mining operations, primarily coal mining, which was first developed by German concessionaires in the early 20th century, provide support for Shandong's industrial foundation. Some of China's greatest coal reserves are located in the southern Shandong region, near Yanzhou and Tengzhou (Shandong - Economy, 2022). These are some of the causes for the rise in the level of PM2.5 in this province.

Jiangsu: China's east coast is home to the province of Jiangsu. The province is bordered to the east by the Yellow Sea, to the southeast by Shanghai Municipality, to the south by Zhejiang, to the west by Anhui, and to the north by Shandong. The provincial capital is Nanjing, which served as both the Nationalist government's (1928-49) and the Ming dynasty's (1368–1644) southern capital of China. Since ancient times, the city has also served as the economic and cultural hub of southern and southeast China. Jiangsu is a key industrial hub and one of China's most economically developed provinces. It consistently produces one of the highest levels of industrial output among the provinces. The processing sector is the most important of the pillar sectors, which also include those that produce electronics, petrochemicals, textiles, foodstuffs, and building materials. Numerous cities, including Changzhou, Nanjing, Nantong, Suzhou, Wuxi, and Yixing, have significant development zones (Jiangsu - Resources and power, 2022). Due to these developments, the quality of air deteriorates in this province.

Zhejiang: South-eastern China's Zhejiang province is amongst the smallest province-level political units in China, yet it is also one of the most densely populated and prosperous. It is a coastal province that is bordered to the east by the East China Sea, to the south by the provinces of Fujian, to the southwest by the province of Jiangxi, to the west by the province of Anhui, to the north by the province of Jiangsu, and to the northeast by Shanghai Municipality. Hangzhou serves as the capital of the province. Light industry accounts for the majority of Zhejiang's wealth. This partially reflects the province's historical status as a major textile producer since the 1890s as well as a commercial and handicraft hub. After 1949, Hangzhou began to develop into a significant industrial hub, producing a wide range of products for both the industrial and consumer markets, including radios, televisions, textiles, machines, and chemicals. A significant industrial hub, Ningbo produces tractors, electronics, and petrochemicals (Zhejiang - Economy, 2022).

3.2. Data

Before getting into the specifics of the methodologies used in this thesis, it is crucial to understand and have an overview of the data that is used in the models. The main objective of this chapter is to discuss the key datasets that are used to create the models and serve as inputs to run them. For this purpose, a total of two datasets are being used. The first dataset offers values that reflect economic growth, and the second dataset includes the concentration of the relevant air pollutant. Each of these dataset's sources, contents, and data wrangling procedures are covered below.

3.2.1 Particulate Matter (PM2.5) data

Information regarding the pollutant in question, in this case fine particulate matter (PM2.5), is provided in the first dataset. A mixture of solid and liquid particles suspended in the air is referred to as particulate matter (PM). These are divided into three groups: coarse, fine, and ultrafine. Larger and relatively heavier than fine particles, coarse particles have a diameter of 2.5 micrometres to 10 micrometres (approximately 25 to 100 times thinner than a human hair), and they have a tendency to settle. Examples include pollen, dust, and spores. PM2.5 refers to particles that have a diameter of less than 2.5 micrometres and are suspended for a prolonged period of time (Important?, 2022). The dataset is produced from two different sources and then combined to produce results. First, the air quality index (AQI) location names for each province/municipality are derived from "China AQI Archive (Feb 2014 – Feb 2016) dataset from the Harvard dataverse website along with their coordinates (**https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/GHOXXO**). Access to more than 2,000 dataverses, 75,000 datasets, and 350,000+ files from institutions, groups, and individuals at Harvard and elsewhere is made possible by Harvard Dataverse (Harvard Dataverse, 2022). There are multiple air monitoring locations set up in these areas which provide the levels of air pollution. Using the station names for each area, daily concentrations of PM2.5 data is then collected from Air Quality Historical Data Platform (**https://aqicn.org/data-platform/register/**). The data is provided in micrograms per cubic metre ($\mu g/m^3$). Founded in 2007, the World Air Quality Index project is a nonprofit endeavour. Its goal is to raise public awareness of air pollution while providing accurate global data on air quality covering more than 30,000 stations for more than 130 countries and 2000 major cities (project, 2022). This platform also provides data for other pollutants such as particulate matter with diameter 10 micrometers and smaller (PM10), ozone (O3), nitrogen dioxide (NO2), sulfur dioxide (SO2) and carbon monoxide (CO).

The process to get the data ready required to include only the PM2.5 pollutant column and remove the other pollutants as they are not being analysed in this thesis. Formats of the date and province column is corrected to reflect the right format. The initial data consisted of the year 2022 but this year is excluded as there isn't sufficient data generated for this period. The daily level of PM2.5 pollutant data is then converted to quarterly data. This is done in order to correlate the level of the relevant pollutant with the quarterly GDP data. Figure 1 depicts the quarterly concentration of PM2.5 for all the selected provinces and municipalities from 2014 – 2021. The legend represents the names of these provinces and municipalities in China. It is observed that the level of PM2.5 for almost all the areas have decreased over the years. While Shandong experienced a spike in the fourth quarter of 2018, Beijing experienced one in the same quarter of 2015. In Chongqing, the concentration of PM2.5 increased in the first quarters of 2014, 2015, and 2016. In both 2014 and 2017, the first quarter saw an increase in Guangdong. In contrast, Hebei experienced levels that rose in the first and fourth quarters of 2014 and 2015, the fourth quarter of 2016, and the first quarter of 2017. Jiangsu experienced a jump in the first quarter of 2014, second quarter of 2017, first and fourth quarters of 2018, and first quarter of 2020 whereas Hubei experienced a spike in the first and fourth quarters of 2016. The fourth quarters of 2014, 2015, and 2017 had high levels in Jilin. In the fourth quarter of 2017 and the second quarter of 2018, Shanghai experienced a surge. In the fourth quarter of 2014, as well as the first and fourth quarters of 2015, Tianjin saw an increase. In addition to the third quarter of 2014, Zhejiang also had a spike in the first quarters of 2014 and 2019. The concentration values for the entire quarterly dataset range from 1.00 $\mu g/m^3$ to 187.00 $\mu g/m^3$, with a mean of 50.86 $\mu g/m^3$ to 207.92 $\mu g/m^3$ and a median of 1.0 $\mu g/m^3$ to 155 $\mu g/m^3$. The Hebei province had the highest quarterly concentration of PM2.5 during the first quarter of 2014. The Shandong province's first and second quarters of 2018 and the Hebei province's first quarter of 2020 and third quarter of 2021 both saw the lowest concentrations.

The original dataset included 972,567 daily PM2.5 concentration measurements from 2014 to 2021, but 6,014 of those daily concentration measurements were missing. After converting it into quarterly data it consisted of 351 quarterly PM2.5 concentration as Hubei was missing PM2.5 observations in the first quarter 2014. This is evident in figure 1 as the observations for the Hubei plot begin in the second quarter of 2014. Few outliers are observed in almost every area's data.

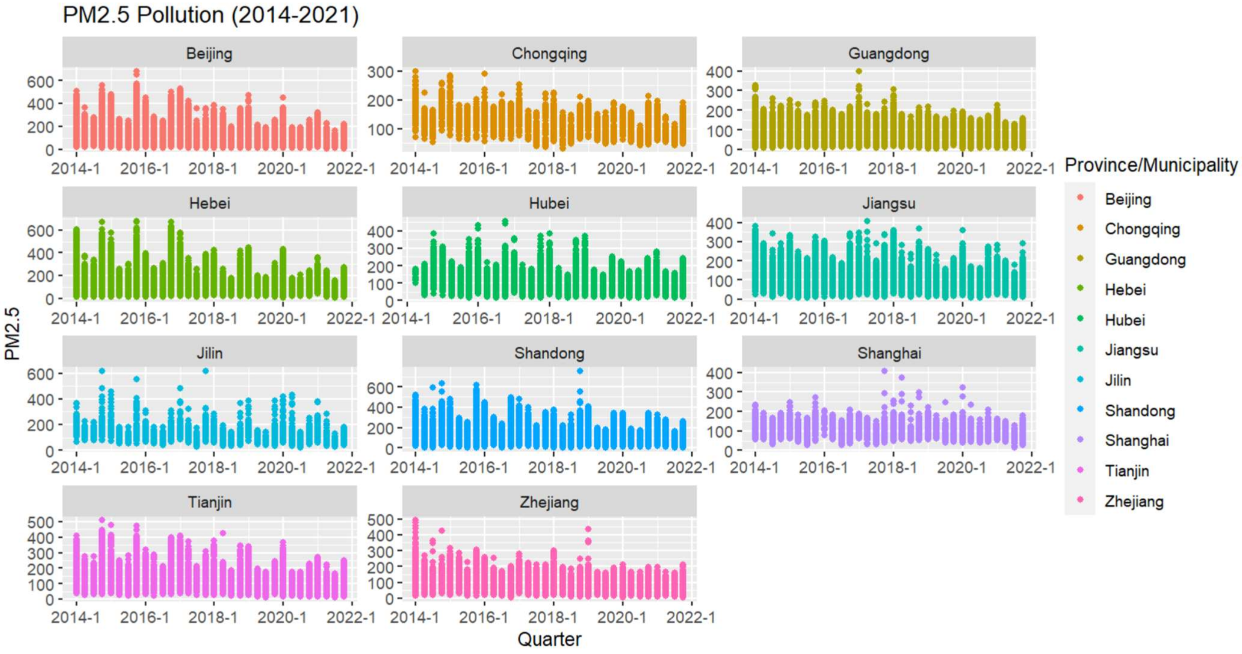

Figure 1: Quarterly PM2.5 pollution from 2014-2021

3.2.2 Gross Domestic Product (GDP) data

The second dataset gives information about the gross domestic product of 7 provinces and all 4 municipalities of China. A measure of an economy's total size and health is its GDP. GDP calculates the total market value (gross) of all domestically produced goods and services in a given year. It tells us whether the economy is growing by creating more goods and services or collapsing by producing less as compared to previous periods (What Is GDP, and Why Is It Important?, 2022). This information for this thesis is derived from EPS China Statistics website (**http://www.epschinadata.com/index.html)**. This website has the most extensive collection of facts on China. It is a platform which offers services like data export, visualisation, processing, analysis, forecasting and data retrieval (LibGuides: Guides for EPS China Data Statistics Databases: Home, 2022). For the eight years between 2014 and 2021 that was chosen for this analysis, quarterly GDP data in yuan (¥) was retrieved for all the specified areas. The formats for the variables, such as the provinces and municipalities, year, and GDP values, was adjusted and converted into the correct formats. This was done in order to prepare the data for the model to ensure accurate computations, if any. The overall dataset consisted of 32 quarters for each province and municipality which made up a total of 352 observations. Table 1 displays the GDP quantiles for each of the chosen areas, including minimum, maximum, average, etc. For instance, Beijing's minimum and maximum GDPs throughout the course of eight years was ¥ 4,412.95 and ¥ 16,103.95 million respectively. The names of Chinese provinces and municipalities are represented in the legend of Figure 2, which shows the quarterly GDP. Over time, it has been seen that the values have risen. The Jilin province reported the lowest GDP of ¥ 2305.10 million in the first quarter of 2014, and the Guangdong province reported the highest GDP of ¥ 124,369.67 million in the fourth quarter of 2021.

Table 1: Quantiles of GDP data for each province/municipality (100 million yuan (¥))

Province	***Count***	***Minimum***	***Median***	***Maximum***	***Sum***	***Mean***
Beijing	32	4,412.95	9,342.60	16,103.95	567,263.41	17,726.98
Chongqing	32	2,982.71	6,217.88	11,231.25	390,191.10	12,193.47
Guangdong	32	13,636.91	28,998.53	51,511.78	1,795,387.19	56,105.85
Hebei	32	5,426.83	11,195.24	20,009.76	657,811.30	20,556.60
Hubei	32	5,137.26	10,967.33	20,159.58	704,767.10	22,023.97
Jiangsu	32	12,892.85	27,996.69	49,892.52	1,711,643.73	53,488.87
Jilin	32	2,305.10	3,956.58	7,381.60	252,370.17	7,886.57
Shandong	32	11,994.96	24,128.83	42,318.77	1,398,995.15	43,718.60
Shanghai	32	5,313.07	10,205.75	17,611.52	618,580.35	19,330.64
Tianjin	32	2,874.35	5,753.93	10,233.30	331,989.65	10,374.68
Zhejiang	32	7,768.46	17,162.49	29,385.43	1,035,576.55	32,361.77

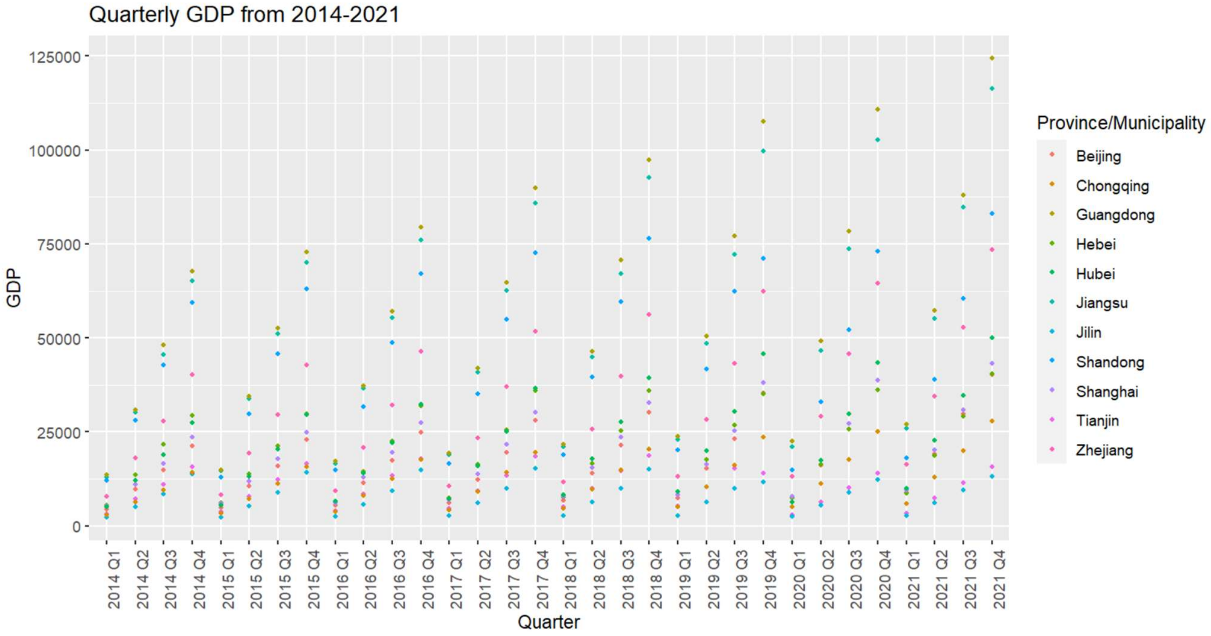

Figure 2: Quarterly GDP from 2014-2021 for each province and municipalities

3.3 Methods

This study adapts two main approaches to arrive at the end result. It follows two frameworks, the first one is introduced in 1986 by Trevor Hastie and Robert Tibshirani who developed the GAM models. The other approach, which is ARIMA was introduced by Box & Jenkins for the first time in 1970 (Swaraj et al., 2021). The GAM model was implemented to analyse the significance of the correlation between GDP and PM2.5 levels. On the other hand, ARIMA model was used to analyse and forecast both of these variables and their patterns. In this section, both the approaches and models and their selection criteria are thoroughly detailed. Additionally, it provides a summary of the underlying framework and its corresponding sources as well as discusses the models' advantages and disadvantages and their assumptions, if any.

3.3.1 Generalised Additive Models (GAM)

The first method used in the analysis is the generalised additive model (GAM). A fundamental and widely used form of predictive analysis is linear regression. The link between one dependent variable and one or more independent variables is explained using these regression estimations. The covariates $X_{1,}X_2, \dots, X_p$are assumed to have a linear (or another parametric) form in likelihood-based regression models like the normal linear regression model and the linear logistic model (Gamma Distribution Explained | What is Gamma Distribution?, 2022). Regression analysis has three main applications: assessing predictor strength, forecasting an effect, and predicting or forecasting trends (What is Linear Regression? - Statistics Solutions, 2022). One of the main assumptions of these models is that the residuals will follow a conditionally normal distribution (General linear model - Wikipedia, 2022). Unfortunately, this isn't true most of the time. Thus, models like GLM and GAM come into the picture. Generalised additive models (GAM) are just another extension of linear models. GAM allows adaptation of modelling non-linear data while preserving explainability. These models replace the linear form $\sum \beta_j X_j$ with a sum of smooth functions $\sum s_j (X_j)$, which is introduced by Trevor Hastie and Robert Tibshirani (Hastie and Tibshirani, 2022). When compared to Generalized Linear Models, such as Linear Regression, a GAM is a linear model with a significant difference. Non-linear feature learning is permitted for GAMs. With GAMs, there is less of a need that the connection be a simple weighted sum and more of an assumption that the result can be represented by the sum of any number of different functions of each feature. To achieve this, a flexible function that supports nonlinear relationships in place of the linear regression's beta coefficients is used. A spline is the name for this adaptable function. Splines are intricate mathematical operations that let us represent non-linear interactions for every feature. A GAM is made up of a lot of splines and the end result is a highly flexible model with some of the linear regression's explainability (What is a Generalised Additive Model?, 2022). The $s_j(.)'s$ here are unknown functions of the covariates $X_{1,}X_2, \dots, X_p$. Any likelihood-based regression model can use the approach (Hastie and Tibshirani, 2022).

In other words, GAM structure can be expressed as:

$$y = \alpha + X\theta + \sum_i f_i\,(x_i) + \sum_j f_j\,(x_{1,j,}x_{2,j}) + \varepsilon$$

$$\text{Where } \mu = E(Y)$$

$$\text{And } Y \sim EF(\mu, \varphi)$$

Y is the dependent variable/responsible variable, $EF(\mu, \varphi)$ denotes the distribution of the exponential family when the location and scale parameters are μ and j respectively. In order to account for non-linear relationships between the covariates and the response variable, f_i and f_j are nonparametric smooth functions. X is a row of the parametric model matrix and θ corresponds with the parameter vector. ε is the normal error term ($\varepsilon \sim N(0, \sigma^2)$) and the i, jth explanatory variables are x_i and x_j respectively. Using smooth functions, the dependence of the response on the covariate is described. These flexible predictors are smooth functions, which can find subtle smooth patterns in the data (Soleimani, Akbari, Saffari and Haghshenas, 2022).

Note that the word "nonparametric" in the context of regression models refers to the fact that the shape of predictor functions is entirely decided by the data, as opposed to parametric functions, which are normally defined by a limited number of parameters. This may enable more adaptable estimation of the underlying predictive patterns without requiring prior knowledge of the patterns' appearance. Both two-dimensional smoothers and parametric terms may be present in GAMs. Moreover, GAM enables a variety of link functions, just like generalised linear models (GLM).

Selection Criteria:

There are three main reasons in selecting this method to determine the relationship between the two variables in this paper (GDP and PM2.5).

1) Interpretability: The first is the interpretability. When a regression model is additive, it is not necessary to know the values of the other variables in the model to evaluate the partial derivative, which measures the marginal impact of a single variable. Therefore, by simply examining the model's output, conclusions can be derived about the influence of the predictive variables that are understandable to non-technical people. Controlling the predictor functions' level of smoothness is another crucial aspect of GAM. Even though the facts at hand might indicate a noisier relationship, prior view that the predictive relationships

are essentially smooth in nature can be imposed. For instance, in the model used for this study, a smooth function was assigned to the PM2.5 variable, which is a predictive variable in order to make it smooth in nature. This is crucial for both the interpretation of the model and the validity of the findings.

2) Flexibility and automation: During model estimation, predictor functions are automatically generated. There is no need to foresee the kinds of functions that will be required. This will not only save time, but it will also enable to identify trends that a parametric model could have missed.
3) Regularisation: As it was indicated earlier, the GAM framework enables the regulation of the predictor functions' smoothness to avoid overfitting. Bias/variance trade-off can be directly addressed by adjusting the wiggliness of the predictor functions. For instance, in our model, the number of knots (k) was assigned to a value of 100 to control the wiggliness of the PM2.5 function. Additionally, there are connections between Bayesian regression and l_2 regularisation and the types of penalties used in GAMs.

Therefore, GAM delivers a regularised and interpretable solution when a model includes nonlinear effects, whereas other approaches typically lack at least one of these three characteristics. In other words, GAMs find a good mix between the extremely flexible "black box" learning methods and the interpretable but biased linear model (LARSEN, 2015).

Advantages and Disadvantages:

When there are many potential predictors, GAM's main benefit is its capacity to model highly complex nonlinear relationships. These models also have the capacity to work with categorical predictors. The amount of flexibility can be used to describe how smoothly the function runs. GAM can be applied to the qualitative data sets as well. In addition to its positives, GAM has various drawbacks as well. The fundamental drawback of GAM is its computational complexity, which also makes it vulnerable to overfitting like other nonparametric approaches (Soleimani, Akbari, Saffari and Haghshenas, 2022). To prevent this, it is a good idea to compare the model fits of a GLM and a GAM and determine whether the additional complexity of GAMs is required to produce a decent fit to the data. It is recommended to employ a GLM model when the fit of a GLM and GAM are comparable (Generalized Additive Model, 2022). The other disadvantages are that the users of GAM must manually input the addition and interaction terms, which can be fitted using splines or smoothers in two dimensions. Additionally, additivity limits the influence of numerous significant interactions in the model (Soleimani, Akbari, Saffari and Haghshenas, 2022).

3.3.2 AutoRegressive Integrated Moving Average (ARIMA) Models

The second method employed for the purpose of forecasting the variables is the ARIMA model. Weakly stable stochastic time series are modelled as two polynomials using an ARMA model, or Autoregressive Moving Average model. These polynomials' first one is for autoregression (AR), and their second one is for the moving average (MA). An ARMA model is a stationary model; in the absence of stationarity, stationarity can be attained by taking a number of differences. This is when ARIMA models are used. Differencing is what sets ARMA and ARIMA models apart. The ARIMA model's "I" stands for integrated (ARMA model, 2022). These models provide another method to time series forecasting (Chapter 8 ARIMA models | Forecasting: Principles and Practice (2nd ed), 2022). A statistical analysis model called an autoregressive integrated moving average, or ARIMA, uses time series data to either better comprehend the data set or forecast future trends. Regression analysis in the form of this model measures the strength of one dependent variable in relation to other varying variables. Instead of using actual values, the model seeks to forecast future values by analysing differences between values in the series (Autoregressive Integrated Moving Average (ARIMA), 2022). Generally, ARIMA models are written as ARIMA (p, d, q), where p is the order of the autoregressive model, d is the level of differencing, and q is the order of the moving-average model. In order to forecast future values from previous data, ARIMA models first transform a non-stationary time series into a stationary one using differencing. To predict future values, these models utilise "auto" correlations and moving averages over residual errors in the data (Understanding ARIMA Models for Machine Learning | Capital One, 2022). Lately, these models are used to forecast PM2.5 concentrations in various studies. For instance, in a study conducted by (Zhang et al., 2018) to forecast the concentration of PM2.5 in Fuzhou, China, an ARIMA model was utilised (Air Pollution PM2.5 Data Analysis in Los Angeles Long Beach with Seasonal ARIMA Model, 2022). Understanding the three terms within the ARIMA model's name is important to understand how the model works:

- AutoRegressive (AR(p)) - Lagged values of y up until the p-th time in the past are used as predictors in the regression model known as autoregressive - AR(p). Here, p = is the model's number of lagged observations, ε is white noise at time t, c is a constant, and φs are parameters.

$$\widehat{y_t} = c + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots + \emptyset_p y_{t-p} + \epsilon_t$$

- Integrated I(d) - Until the original series becomes stationary, the difference is multiplied by d times. A time series is considered stationary if its characteristics are independent of the observational time.

  $By_t = y_{t-1}$ where B is referred to as a backshift operator

  Thus, a first order difference is written as

  $$y'_t = y_t - y_{t-1} = (1 - B)y_t$$

  In general, a dth order difference can be written as

  $$y'_t = (1 - B)^d y_t$$

- Moving average MA(q) - A regression-like model on prior forecasting errors is used in a moving average model. Here, c is a constant, θs are parameters, and ε is white noise at time t.

$$\widehat{y_t} = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

The resulting ARIMA(p,d,q) model is obtained by combining all three of the above model types.

$$\widehat{y'_t} = c + \emptyset_1 y'_{t-1} + \emptyset_2 y'_{t-2} + \cdots + \emptyset_p y'_{t-p} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

(Understanding ARIMA Models for Machine Learning | Capital One, 2022)

Selection Criteria:

This study's variables were predicted using an ARIMA model is because these models are typically employed when there is a regular interval of time series data such as daily, monthly, quarterly etc (ARIMA Modeling, 2022). This is relevant to this investigation because both variables' observations are made quarterly. The ability to forecast future values using previous data for PM2.5 and GDP was another factor in the choice of this approach. Last but not least, it is well known that these models are typically used when the data exhibits signs of non-stationarity (Chao, 2022). This holds true for the two datasets in this study. As a result, it was decided that these models would provide better predictions for the datasets that are available.

Assumptions:

- The foundation of ARIMA models is the idea that past values may still have an impact on present or future values. For example, in this study after analysing the significance of PM2.5 on GDP, the prediction of GDP values is made on the assumption that level of PM2.5 has an influence on these values. Although in many cases this assumption will be accurate, it is not always the case (Autoregressive Integrated Moving Average (ARIMA), 2022).
- Data should be steady, which means that its characteristics shouldn't change depending on when it was recorded. A series that exhibits cyclic behaviour and white noise can alternatively be categorised as a stationary series (Chatterjee, 2022).
- Since ARIMA only considers one variable, the data must be univariate. Auto-regression focuses exclusively on regression using historical values (Chatterjee, 2022).

Advantages and Disadvantages:

While these time series models have some benefits, they have some limitations too. ARIMA is a method that is linear, making it useful and widely used in the field for testing, interpreting the data, and establishing baseline forecasting scores. These models can perform noticeably better if calibrated properly with lagged values (p, d, and q). Its algorithm is among the top choices among analysts and data scientists due to its simplicity and explainability. Working with ARIMA at scale is not without its advantages and disadvantages, though. Some of the advantages of these models are that they are easy to comprehend and interpret. The one thing that is always beneficial is the simplicity and interpretability of the model. It has fewer hyperparameters meaning fewer variables, making it simpler to maintain the configuration file if the model is put into use. It is also useful when modelling non-stationary data (Understanding ARIMA Models for Machine Learning | Capital One, 2022). Some of the limitations of these models are the exponential time complexity. If p and q are high, there are more coefficients to fit as p and q increase, which multiplies the time complexity. Due to its difficulty in implementation, these algorithms have Data Scientists looking towards Prophet and other techniques. However, it also relies on how complex the dataset is. It's possible that the data is too complex and there isn't an ideal answer for p and q. The failure of ARIMA is quite unlikely, but there's still a chance of that happening. The algorithm requires a large amount of data to operate, especially if the data is seasonal. For Short Life-Cycle Products, for instance, using three years of historical demand is probably insufficient for a reliable estimate (Bajaj, 2022). Other drawbacks of this approach include its poor performance for long-term forecasts. Additionally, seasonal time series cannot be used with it (Understanding ARIMA Models for Machine Learning | Capital One, 2022).

4. Application and Results

4.1 Application and results of Gamma GAM model

This chapter demonstrates the practical application of the theoretical framework presented in chapter 3 (3.3) to the construction of a real model. It first explains how the model can be used to fit the scenario for this study, and then it talks about the results of those models. It should be noted that although an ARIMA model is employed in this chapter to estimate GDP values and PM2.5, the model can also be used to estimate other pollutants or factors. The input data to the model as well as the settings in each component can be changed to achieve this. However, the information must be a regular time series. Finally, all of the analysis was carried out using the R software after installing the necessary packages (refer appendix for the codes).

First, a linear regression model (lm()) was fitted to the quarterly data, which consisted of PM2.5 and GDP levels for the chosen regions of China, to test the linearity of the two variables (refer section 3.1). Here, the GDP values (Y_i) serve as the response variable, and the the PM2.5 concentrations (x_i) serve as the independent or the covariates or explanatory variable. The model's output shows that the data are non-linear and have a negative correlation (refer to Figure 6 in appendix). As a result, we choose to use the "mgcv" package to fit a GAM model. To assess the model's viability, many GAM models with various conditional response distributions and link functions are developed. To fit the model formula, a smooth function using s() and the number of knots (k) are supplied, along with the kind of basis function (bs). Gamma distribution with log link function was selected after fitting numerous GAM models with various conditional distributions, including gaussian, poisson, quasipoisson, and negative binomial distributions with their relevant link functions. In many scientific disciplines, continuous variables with skewed distributions that are always positive are modelled using the gamma distribution, a continuous probability distribution. When the times between events matter, it naturally occurs in processes (Gamma Distribution Explained | What is Gamma Distribution?, 2022). This conditional distribution seems to fit well because the GDP variable is a continuous data set with a positively skewed distribution (see the histogram plot in the appendix). This can be displayed as follows for the positive continuous GDP data (Y_i):

$$Y_i \sim Gamma(\mu_i, \sigma^2)$$

$$\eta_i = \log(\mu_i) = \beta_0 + f_1(x_i) + f_2(x_i) + \cdots + f_{11}(x_i)$$

Where Y_i represents the GDP values, the level of PM2.5 is represented by x_i and $f_1(.)$, $f_2(.)$…, $f_{11}(.)$ are functions of x_i for each province/municipality. Log link is used in this model as the values are required to be in positive. If the inverse or identity link were employed, nothing would require μ_i to be positive and there may be a potential that predictions for negative values would be made (MTHM506 Statistical Data Modelling Topic 3-Another example of GAMs, 2022). With the number of knots set to 30 and the basis function being cubic splines (cs), this Gamma model was fitted using the smooth function on PM2.5. By utilising the "by=" option and selecting the limited maximum likelihood (REML) method, an interaction between the smooth function (PM2.5) and factors (province/municipalities) is introduced. The parametric coefficients and the approximate significance of the smooth terms make up the model's output. For easier visualisation, these are presented in two different tables. The parametric coefficients are shown in Table 2, which has five columns. The model's parametric coefficients are shown in the first column. Its estimations are shown in the second column, followed by its standard error, t-test result, and lastly the p-values. This table shows that the model's intercept (GDP) is significant because the p-value is less than 0.05. The approximate significance results for the smooth terms are shown in Table 3. The first column of this table's five columns contains the smooth terms for each province and municipality. The complexity of the smooth terms is represented by the second column's effective degrees of freedom (e.d.f.). The higher the edf value, the wigglier the curves are. The test statistics used in an Anova test to determine the overall significance of the smooth are shown in the following two columns, Ref.df and F. The test's outcome, or approximate p-value, is shown in the final column. Other than the province of Hebei, it has been noted that all other locations are significant. Despite the fact that the fixed effects for the Hubei and Shanghai provinces and Beijing Municipality are only significant at the 0.001 and 0.01 levels, respectively.

Table 2: Parametric Coefficients of Gamma GAM model

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.76877	0.04085	239.1	<2e-16	***

Table 3: Approximate significance of the smooth terms for Gamma GAM model

Smooth terms	*edf*	*Ref.df*	*F*	*p-value*	
s(PM25): Province Beijing	1.0328	28	0.185	0.01761	*
s(PM25): Province Chongqing	1.12549	25	0.464	0.000496	***
s(PM25): Province Guangdong	3.56719	19	5.697	< 2e-16	***
s(PM25): Province Hebei	0.01375	27	0	0.4271	
s(PM25): Province Hubei	2.33103	28	0.363	0.005449	**
s(PM25): Province Jiangsu	9.27063	28	3.522	< 2e-16	***
s(PM25): Province Jilin	11.29934	27	2.38	< 2e-16	***
s(PM25): Province Shandong	12.86362	27	2.392	< 2e-16	***
s(PM25): Province Shanghai	1.02818	24	0.345	0.002659	**
s(PM25): Province Tianjin	2.02825	26	0.58	2.92E-04	***
s(PM25): Province Zhejiang	2.34365	25	1.282	3.82E-07	***

The Gamma GAM model, as a whole, explains 59.4% of the model's deviance, and its adjusted R-squared value is 0.532. The model's adjusted R-squared value might rise if more independent or explanatory variables are included.

Model and residual checking:

By examining its residual plots and utilising "gam.check()" to do model checking, this model's validity is examined. This process results in two things. 1) Residual plots (figure 3), and 2) a summary of the model's smoothed functions (table 6 refer appendix). Figure 3 shows the Q-Q plot, which is the first figure (top left), and shows that the bulk of residuals, with the exception of a few at each end, sit on the diagonal line, suggesting that the Q-Q plot appears to fit the model well. The residuals vs. fitted values plot in the second (top right) plot does not display a pattern which is favorable. The function is flexible enough to capture the peaks, which explains this. To establish whether the residuals follow a N (0, 1) distribution, look at the residual's histogram in the third (bottom left) plot. This plot provides information that is similar to that provided by the Q-Q plot, hence it may be said to be somewhat repetitive. The response (Y_i) is shown against the fitted/predicted values $(\hat{Y}_i)$ in the fourth (bottom right) plot. The points should lie or scatter evenly on the diagonal line if the model fits well, however as was already noted, the plot shows a pattern. This figure's information is identical to what is gained from the residuals vs. fitted values plot.

To determine whether the smooth functions' flexibility is sufficient, the second output, or the summary of the smoothed functions table, is then evaluated. Table 6's (refer the appendix) first column, labelled k′, displays the number of parameters this function contains. The intercept of a rank q will cause q-1 to be displayed. Since q = 30 in our model, k′ equals 29. The second column's edf provides details on "how many parameters from the available k′ did it use to estimate the function after penalization. Despite the model being penalised for not overfitting the data, rank 30 was chosen, resulting in a range for each position of around 0.01 to 13 degrees of freedom. If a function's edf and k′ differ by less than one, it can require more degrees of freedom (i.e., fewer than one parameter). The difference between k' and the edf in this case is 0.94, or around 1. As a result, the functions rank q was not increased. The k-index and associated p-values are included in the third and fourth columns, respectively. Once more, these are useful in determining whether or not the function should be given more flexibility. They are only used as indicators. If the k-index is less than 1 or the p-value is extremely small, the k' is too low. Results show that the k-index and p-values in the chosen GAM model are not exceptionally low, so it is not necessary to increase the rank of the function. It can be judged that this function is adequate despite the fact that these tests are not official.

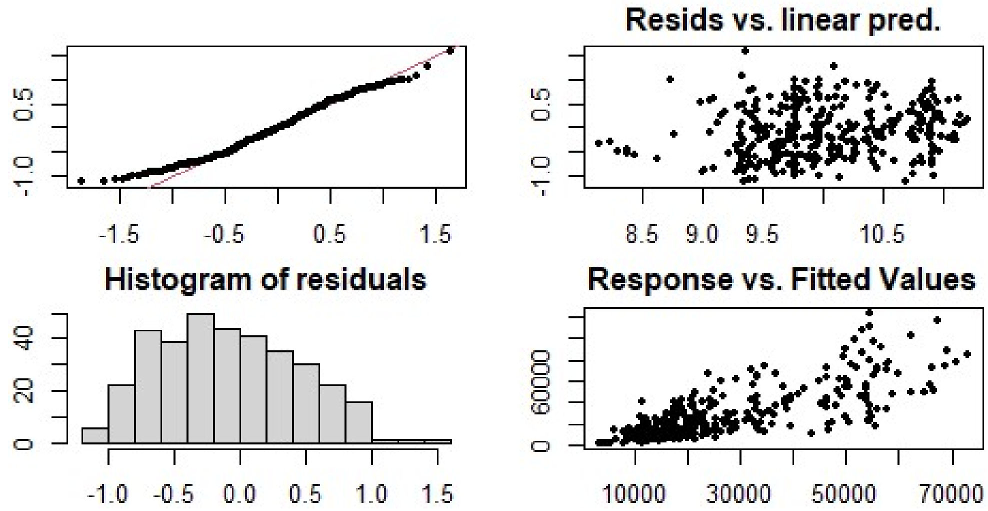

Figure 3: Residual plots for Gamma GAM model

Model comparison:

To compare the different GAM model, akaike information criterion (AIC) value is examined by using “AIC()”. Gamma GAM model’s AIC is the lowest at 7607.560 when compared with other GAM models (Gaussian = 7817.353, negative binomial = 22521.691).

Examining the output of the GAM model reveals a substantial and inverse relationship between GDP and PM2.5. However, when broken down by province, it becomes clear that, with the exception of Hebei, all other provinces exhibit significant values. This investigation establishes that PM2.5 levels have an impact on GDP figures. These findings concur with those made by (Hao et al., 2018) in his analysis of the effects of air pollution on economic growth (refer section 2.3 for more details). If there were more factors that affected the fluctuation of GDP values in addition to the amount of PM2.5, more weightage and more effective results could have been generated.

4.2 Application and results of ARIMA time-series model

The next step is to forecast the GDP values and PM2.5 levels in the chosen provinces and municipalities of China after validating the relevance of the association between GDP and PM2.5. The R software is used to fit the ARIMA model in order to achieve this (refer 3.3.2 for details of the framework). This is accomplished by using "ts()" to turn the variable data into a time series and setting the frequency to 4 since the data is quarterly. Once the data has been transformed into a time series, it is examined to see if the underlying trend is stationary and whether there are any indications of seasonality. This was accomplished by displaying it for visual purposes (see figures 7 and 9 in the appendix), as well as by plotting the ACF and PACF of the data (refer figure 8 & 10 in appendix). The figure, like the ACF and PACF plots, indicated a trend but not seasonality. Thus, the trend was eliminated by applying the first set of differencing. The plot that resulted appeared much more stationary than the initial one. This was verified by plotting ACF and PACF once more (refer figure 8 & 10 in appendix). These charts make it unclear in what order the ARIMA model should be fitted. Therefore, using "arima()" in the forecast package, we fitted 5 models, each with a different order for the two variables. Model 3 (ARIMA(1,1,1)) is chosen for PM2.5, while Model 4 (ARIMA(2,1,0)) is chosen for GDP, out of these 5 models. The results of the chosen ARIMA models for both variables are displayed in Tables 4 and 5. Table 4's coefficient values are examined to make sure they are not too close to 0 because this would make them redundant. It can be shown that both coefficient values are substantial enough to be kept in the model in this instance.

Table 4: Output of ARIMA model for PM2.5 (ARIMA (1,1,1))

Call:		
arima(x = qrtly.ts, order = c(1, 1, 1))		

Coefficients:		
	ar1	ma1
	-0.3208	-0.5907
s.e.	0.2109	0.1786
sigma^2 estimated as 544.8: log likelihood = -142.09, aic = 290.17		

The same holds true for table 5. The output of the chosen ARIMA model for GDP has all statistically significant coefficients with order (2,1,0).

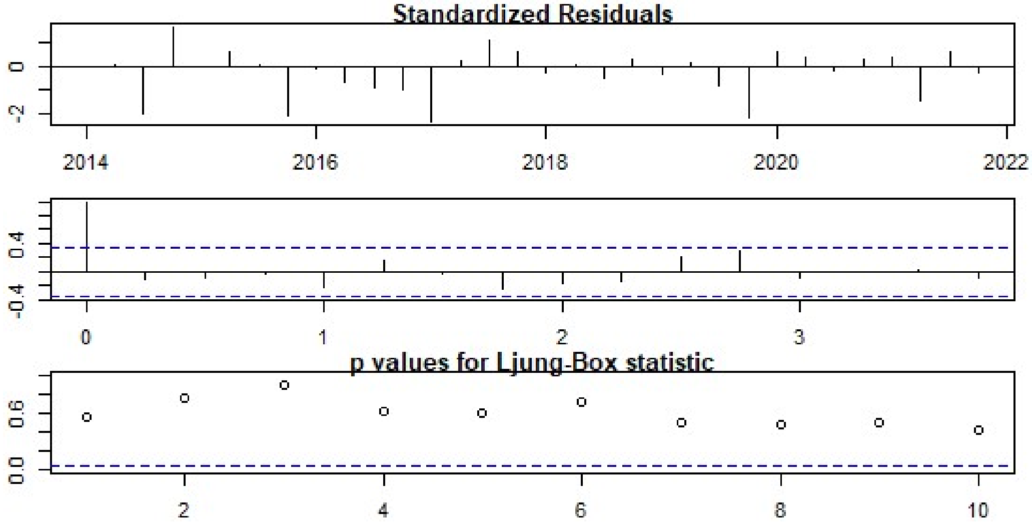

Figure 4: Residual plot for PM2.5 ARIMA model 3

After that, the residual plots were examined to determine whether the models were reliable. Figure 4 shows that there are no evident correlations in the residuals for the PM2.5 model (top plot). P-values are high (bottom plot), and the ACF is not very evident (middle plot).

Table 5: Output of ARIMA model for GDP (ARIMA(2,1,0))

Call:		
arima(x = gdp_qrtly.ts, order = c(2, 1, 0))		
Coefficients:		
	ar1	ar2
	-1.1393	-0.6552
s.e.	0.1303	0.1264
sigma^2 estimated as 90671111: log likelihood = -328.87, aic = 663.74		

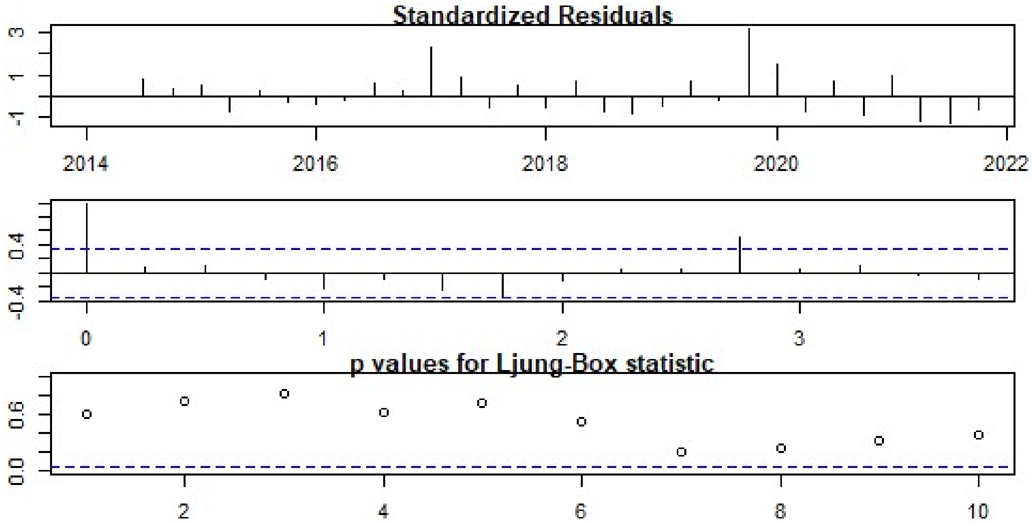

Figure 5: Residual plot for GDP ARIMA model 4

Figure 5 displays similar outcomes. No clear correlation can be seen in the residuals (top plot), and the ACF is not evident (middle plot). Fir this model, the p-values are high enough and not too low.

After reviewing all of the residual plots and model results, both of these models appear to provide a better fit. All the AIC values and residual diagnostics from all the models are compared to further support this idea. The lowest AIC among the other models is found in ARIMA model 3 for PM2.5 with order (1,1,1) and ARIMA model 4 for GDP with order (2,1,0), with values of 290.17 and 663.74, respectively. The usage of "auto.arima()" to automatically estimate the ordering is the last step in ensuring the accuracy of the models. For the PM2.5 model, the auto.arima function calculated an ARIMA(0,1,1) order with an AIC value of 290.23. The estimated model's AIC values are higher than those of our model, which also has an order that is close to the estimated order by this function. On the other side, the order we selected for our GDP model was identical to the estimated order for the GDP model by this function (2,1,0). Its AIC value coincided with the AIC value of our model. The chosen models are used to forecast the variables after the model's orders have been verified. 50 values are provided for the prediction of both variables using the "Predict()" function. The values in the available data are pretty similar to the predicted outputs (refer to table 7 & 8 in appendix). From the predicted values, it can be seen that GDP values will rise and PM2.5 levels will fall during the following few years. This further supports the notion that there is an inverse relationship between PM2.5 and GDP. The results of the study conducted by (Zhang et al., 2018), where the level of PM2.5 also decreased, are similar to the predicted result for PM2.5 level in this study. Hence, we can conclude that with the help of ARIMA models, GDP and PM2.5 values are forecasted from the first quarter of 2022 till the second quarter of 2034. (For the purpose of visualisation refer figures 11 & 12 in appendix which represent these predicted values for both the variables.)

The time limitations to finish this thesis impair the accuracy and complexity of the models. The key ideas that could improve and strengthen the model are outlined in the following paragraphs. First, data on the GDP and PM2.5 for the chosen provinces and municipalities from 2014 through 2021 are accessible. But this thesis primarily examines a broad perspective. particularly in terms of ARIMA time-series modelling. If given more time, the same techniques can be used to do this analysis at the provincial level. The coordinates of the chosen locations are also included in the initial data. A spatio-temporal analysis could be performed using these coordinates. This would be useful for comparing the GDP figures and PM2.5 levels over time and in different locations. This would aid in improving one's understanding and ability to visualise the study. Last but not least, because the original PM2.5 data are daily, seasonality might be incorporated by performing a time-series analysis on the daily data, such as dynamic linear models (DLM). The addition of these additional methods or the broadening and deepening of the study could have an impact on the effects of PM2.5 on GDP and the accuracy of the estimates. As was already mentioned, there may be chance to boost the dissertation's credibility and impact by making the models more complex and broadening the study's scope while also giving the additional time to finish.

5. Conclusion

This dissertation presents the impact that the air pollution has, in this case particulate matter with a diameter smaller than 2.5 micrometres (PM2.5) on the economy of China. Here the growth of the economy is represented by GDP values. Two methods are utilised in this study to carry out this analysis. 1) In the first method regression analysis is used, specifically Generalised Additive Models (GAM) to evaluate the significance and the correlation between the two variables. To do this quarterly data of GDP as well as PM2.5 of 7 provinces and 4 municipalities of China is derived from 2014 to 2021. PM2.5 was originally a daily data but for the purpose of analysis it is converted to quarterly data. A Gamma GAM model with a log link function is fit to the dataset with a smooth function, where appropriate knots and basic function along with the method is specified. This model is also categorised by the provinces selected. Next, a residual check is done along with model comparison where the model with the lowest AIC value is selected. 2) In the second method, a time series analysis is done with the help of ARIMA models. This is carried out for the purpose of identifying a trend in the data and then forecasting the values of both the variables. Two ARIMA models are fit, one for each variable after introducing first differencing in the models. This is done in order to make the underlying trend stationary. After fitting multiple models with different orders, ARIMA model with the order (1,1,1) is selected for predicting the PM2.5 level and an ARIMA model with the order of (2,1,0) is selected for predicting the GDP values. The validity of these models is then checked by performing residual diagnostics and model comparison is done by looking at the lowest AIC value. To further confirm the selection of models, "auto.arima()" function is used to automatically calculate the orders. Once the models are finalised, prediction of the two variables is done from the first quarter of 2022 till the fourth quarter of 2034.

The output from the GAM model suggests that there is a significant relation between the GDP and PM2.5 and that it is a negative correlation. This means that when the levels of PM2.5 decreases, the GDP values increases and vice versa. From the output of the GAM model, it is noticed that apart from all the selected provinces and municipalities of China, Hebei province is not significant. Its p-value is more than 0.1. However, from the observations of GAM analysis, it can be said that PM2.5 does have an impact on the fluctuation of GDP values, which represents the fluctuation of the economy. Next, from the ARIMA model, a trend of these variables is identified. The prediction values derived from the ARIMA models for both these variables suggests that in the beginning of the year 2023 (Q1) the level of PM2.5 would decrease to 114.807 $\mu g/m^3$ overall for the selected locations in China, while the GDP value would increase to ¥ 20,909.83 million. This again confirms the finding that there is a negative correlation between the two variables. There could be many reasons for the decrease of PM2.5 level in most parts of China but the main reason could be the implementation of multiple air pollution control policies that the government has undertaken recently. Clearly the result of controlling air pollution, specifically PM2.5 has shown its impact not only on the health factor but also on the economy of China.

To fully utilise the potential of both of the methodologies utilised in this study, further research on the subject is still needed, as well as a more precise data. For instance, more explanatory variables can be introduced that effect the GDP values in addition to the PM2.5 levels in the GAM models. For example, the total population, total health expenditure of the selected locations etc. This would not only increase the complexity of the model but better explain the model and the fluctuation in GDP values. Another parameter that could be added in the GDP per capita. This would make it easier to explain the fluctuations as it would be in percentages. The other aspect that could be introduced in this analysis other than the economical aspect is health. As PM2.5 has a major impact on the health of the public, this could be another area to look into along with the implementations of air pollution policies and the changes it's brought arounds. Finally, instead of quarterly data, if the annual datasets were available for both the variables over a good period of time, it would be a much accurate analysis. Nevertheless, the observation from the current GAM model and the prediction values from the ARIMA models is still decent enough to prove that PM2.5 has an influence on the growth of the economy.

Bibliography

2021. *Empirical Analysis: Definition, Characteristics and Stages*. [online] Available at: <https://www.indeed.com/career-advice/career-development/empirical-analysis> [Accessed 16 September 2022].

Azam, M., 2016. Does environmental degradation shackle economic growth? A panel data investigation on 11 Asian countries. *Renewable and Sustainable Energy Reviews*, [online] 65, pp.175-182. Available at: <https://www.sciencedirect.com/science/article/pii/S136403211630315X>.

Bajaj, A., 2022. *ARIMA & SARIMA: Real-World Time Series Forecasting - neptune.ai*. [online] neptune.ai. Available at: <https://neptune.ai/blog/arima-sarima-real-world-time-series-forecasting-guide> [Accessed 16 September 2022].

BCCVL. 2022. *Generalized Additive Model*. [online] Available at: <https://support.bccvl.org.au/support/solutions/articles/6000083208-generalized-additive-model#header-page1> [Accessed 16 September 2022].

Briefing, C., 2022. *Beijing City Profile - Industry, Economics, and Policy*. [online] China Briefing News. Available at: <https://www.china-briefing.com/news/beijing-industry-economics-policy/#:~:text=With%20a%20compound%20annual%20GDP,level%20registered%20by%20developed%20countries.> [Accessed 16 September 2022].

Capital One. 2022. *Understanding ARIMA Models for Machine Learning | Capital One*. [online] Available at: <https://www.capitalone.com/tech/machine-learning/understanding-arima-models/> [Accessed 16 September 2022].

Chao, D., 2022. *Time Series Analysis using Arima Model - Analytics Vidhya*. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/11/performing-time-series-analysis-using-arima-model-in-r/> [Accessed 16 September 2022].

Chatterjee, S., 2022. *Time Series Analysis Using ARIMA Model In R | DataScience+*. [online] Datascienceplus.com. Available at: <https://datascienceplus.com/time-series-analysis-using-arima-model-in-r/#:~:text=Assumptions%20of%20ARIMA%20model&text=A%20white%20noise%20series%20and,regression%20with%20the%20past%20values.> [Accessed 16 September 2022].

Chen, V., Deng, W., Yang, T. and Matthews, S., 2012. Geographically Weighted Quantile Regression (GWQR): An Application to U.S. Mortality Data. Geographical Analysis, [online] 44(2), pp.134-150. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4204738/#:~:text=Like%20other%20analytic%20methods%2C%20GWR,2009%3B%20Boots%20and%20Okabe%202007%3B>.

Chetty, P., 2022. *What are the advantages and disadvantages of an empirical study?*. [online] Knowledge Tank. Available at: <https://www.projectguru.in/what-are-the-advantages-and-disadvantages-of-an-empirical-study/#:~:text=Since%20an%20empirical%20study%20contributes,they%20are%20flexible%20to%20incorporate.> [Accessed 16 September 2022].

Chrome.google.com. n.d. *An Introduction to Instrumental Variables*. [online] Available at: <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://mchp-appserv.cpe.umanitoba.ca/supp/mchp/protocol/media/Instrumental_variables.pdf> [Accessed 16 September 2022].

CORP-MIDS1 (MDS). 2022. *ARIMA Modeling*. [online] Available at: <https://www.mastersindatascience.org/learning/statistics-data-science/what-is-arima-modeling/#:~:text=The%20model%20is%20used%20to,daily%2C%20weekly%20or%20monthly%20periods.> [Accessed 16 September 2022].

Ding, Y., Zhang, M., Chen, S., Wang, W. and Nie, R., 2019. The environmental Kuznets curve for PM2.5 pollution in Beijing-Tianjin-Hebei region of China: A spatial panel data approach. *Journal of Cleaner Production*, 220, pp.984-994.

Dong, D., Xu, B., Shen, N. and He, Q., 2021. The Adverse Impact of Air Pollution on China's Economic Growth. *Sustainability*, 13(16), p.9056.

En.wikipedia.org. 2022. *General linear model - Wikipedia*. [online] Available at: <https://en.wikipedia.org/wiki/General_linear_model> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Beijing - People*. [online] Available at: <https://www.britannica.com/place/Beijing/People> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Chongqing - Economy*. [online] Available at: <https://www.britannica.com/place/Chongqing/Economy> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Guangdong - Settlement patterns*. [online] Available at: <https://www.britannica.com/place/Guangdong/Settlement-patterns#ref71343> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Hebei - Climate*. [online] Available at: <https://www.britannica.com/place/Hebei/Climate#ref71097> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Hubei - Resources and power*. [online] Available at: <https://www.britannica.com/place/Hubei/Resources-and-power> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Jiangsu - Resources and power*. [online] Available at: <https://www.britannica.com/place/Jiangsu/Resources-and-power> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Jilin - People*. [online] Available at: <https://www.britannica.com/place/Jilin-province-China/People#ref71050> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Shandong - Economy*. [online] Available at: <https://www.britannica.com/place/Shandong-province-China/Economy> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Shanghai - Economy*. [online] Available at: <https://www.britannica.com/place/Shanghai/Economy> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Tianjin | History, Map, Population, & Facts*. [online] Available at: <https://www.britannica.com/place/Tianjin-China> [Accessed 16 September 2022].

Encyclopedia Britannica. 2022. *Zhejiang - Economy*. [online] Available at: <https://www.britannica.com/place/Zhejiang/Economy> [Accessed 16 September 2022].

Federal Reserve bank of St. Louis. 2022. *What Is GDP, and Why Is It Important?*. [online] Available at: <https://www.stlouisfed.org/open-vault/2019/march/what-is-gdp-why-important#:~:text=GDP%20measures%20the%20total%20market,contracting%20due%20to%20less%20output.> [Accessed 16 September 2022].

Fotheringham, A., Brunsdon, C. and Charlton, M., 2010. Geographically weighted regression. Chichester: Wiley.

Frost, J., 2022. When Should I Use Regression Analysis?. [online] Statistics By Jim. Available at: <https://statisticsbyjim.com/regression/when-use-regression-analysis/> [Accessed 11 September 2022].

Gov.scot. 2022. *Computable General Equilibrium modelling: introduction*. [online] Available at: <https://www.gov.scot/publications/cge-modelling-introduction/> [Accessed 16 September 2022].

Guides.library.illinois.edu. 2022. *LibGuides: Guides for EPS China Data Statistics Databases: Home*. [online] Available at: <https://guides.library.illinois.edu/c.php?g=1059328#:~:text=The%20EPS%20China%20Statistics%20presents,visualization%20display%20and%20data%20export.> [Accessed 16 September 2022].

Hao, Y., Peng, H., Temulun, T., Liu, L., Mao, J., Lu, Z. and Chen, H., 2018. How harmful is air pollution to economic development? New evidence from PM2.5 concentrations of Chinese cities. Journal of Cleaner Production, [online] 172, pp.743-757. Available at: <https://www.sciencedirect.com/science/article/pii/S0959652617325039>.

Hao, Y., Peng, H., Temulun, T., Liu, L., Mao, J., Lu, Z. and Chen, H., 2018. How harmful is air pollution to economic development? New evidence from PM2.5 concentrations of Chinese cities. *Journal of Cleaner Production*, 172, pp.743-757.

Harvard Library. 2022. *Harvard Dataverse*. [online] Available at: <https://library.harvard.edu/services-tools/harvard-dataverse#:~:text=Harvard%20Dataverse%20provides%20access%20to,individuals%20at%20Harvard%20and%20beyond.> [Accessed 16 September 2022].

Hastie, T. and Tibshirani, R., 2022. *Generalized Additive Models*. [online] Pdodds.w3.uvm.edu. Available at: <https://pdodds.w3.uvm.edu/files/papers/others/1986/hastie1986a.pdf> [Accessed 16 September 2022].

Health Organization, W., 2016. Ambient air pollution: a global assessment of exposure and burden of disease. Clean Air Journal, [online] 26(2), p.6. Available at: <https://apps.who.int/iris/handle/10665/250141?locale-attribute=ar&mbid=synd_yahoolife>.

Health.ny.gov. 2022. Fine Particles (PM 2.5) Questions and Answers. [online] Available at: <https://www.health.ny.gov/environmental/indoors/air/pmq_a.htm#:~:text=Fine%20particulate%20matter%20(PM2.5,hazy%20when%20levels%20are%20elevated.> [Accessed 11 September 2022].

Hong-Wei, Y., Toshihiko, M. and Yue, W., 2006. Health and economic impacts of air pollution in China: A comparison of the general equilibrium approach and human capital approach. *Biomedical and environmental sciences : BES*, [online] 18, pp.427-41. Available at: <https://www.researchgate.net/publication/7233425_Health_and_economic_impacts_of_air_pollution_in_China_A_comparison_of_the_general_equilibrium_approach_and_human_capital_approach/citation/download> [Accessed 16 September 2022].

Ieeexplore.ieee.org. 2022. *Air Pollution PM2.5 Data Analysis in Los Angeles Long Beach with Seasonal ARIMA Model*. [online] Available at: <https://ieeexplore.ieee.org/abstract/document/5367074> [Accessed 16 September 2022].

Important?, W., 2022. *What Is PM2.5 and Why Is It Important?*. [online] airveda. Available at: <https://www.airveda.com/blog/what-is-pm2-5-and-why-is-it-important> [Accessed 16 September 2022].

Investopedia. 2022. *Autoregressive Integrated Moving Average (ARIMA)*. [online] Available at: <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp> [Accessed 16 September 2022].

Investopedia. 2022. *Autoregressive Integrated Moving Average (ARIMA)*. [online] Available at: <https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp> [Accessed 16 September 2022].

Iqair.com. 2022. *China Air Quality Index (AQI) and Air Pollution information | IQAir*. [online] Available at: <https://www.iqair.com/us/china> [Accessed 16 September 2022].

Iqair.com. 2022. *Chongqing Air Quality Index (AQI) and China Air Pollution | IQAir*. [online] Available at: <https://www.iqair.com/china/chongqing> [Accessed 16 September 2022].

Iqair.com. 2022. *Guangzhou Air Quality Index (AQI) and China Air Pollution | IQAir*. [online] Available at: <https://www.iqair.com/china/guangdong/guangzhou> [Accessed 16 September 2022].

Iqair.com. 2022. *Shanghai Air Quality Index (AQI) and China Air Pollution | IQAir*. [online] Available at: <https://www.iqair.com/china/shanghai> [Accessed 16 September 2022].

Iqair.com. 2022. *Tianjin Air Quality Index (AQI) and China Air Pollution | IQAir*. [online] Available at: <https://www.iqair.com/au/china/tianjin> [Accessed 16 September 2022].

Iqair.com. 2022. *Wuhan Air Quality Index (AQI) and China Air Pollution | IQAir*. [online] Available at: <https://www.iqair.com/china/hubei/wuhan> [Accessed 16 September 2022].

Jin, Y., Andersson, H. and Zhang, S., 2016. Air Pollution Control Policies in China: A Retrospective and Prospects. International Journal of Environmental Research and Public Health, [online] 13(12), p.1219. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5201360/>.

Jin, Y., Andersson, H. and Zhang, S., 2016. Air Pollution Control Policies in China: A Retrospective and Prospects. International Journal of Environmental Research and Public Health, [online] 13(12), p.1219. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5201360/>.

LARSEN, K., 2015. *GAM: The Predictive Modeling Silver Bullet*. [online] Multi Threaded. Available at: <https://multithreaded.stitchfix.com/blog/2015/07/30/gam/#:~:text=Generalized%20additive%20models%20were%20originally,can%20be%20linear%20or%20nonlinear.> [Accessed 16 September 2022].

Liao, Q., Jin, W., Tao, Y., Qu, J., Li, Y. and Niu, Y., 2020. Health and Economic Loss Assessment of PM2.5 Pollution during 2015–2017 in Gansu Province, China. International Journal of Environmental Research and Public Health, [online] 17(9), p.3253. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7246598/>.

Lin, Y., Zou, J., Yang, W. and Li, C., 2018. A Review of Recent Advances in Research on PM2.5 in China. *International Journal of Environmental Research and Public Health*, 15(3), p.438.

Medium. 2022. *What is a Generalised Additive Model?*. [online] Available at: <https://towardsdatascience.com/generalised-additive-models-6dfbedf1350a#c407> [Accessed 16 September 2022].

Mur, J. and Angulo, A., 2006. The Spatial Durbin Model and the Common Factor Tests. *Spatial Economic Analysis*, 1(2), pp.207-226.

Otexts.com. 2022. *Chapter 8 ARIMA models | Forecasting: Principles and Practice (2nd ed)*. [online] Available at: <https://otexts.com/fpp2/arima.html> [Accessed 16 September 2022].

Pdodds.w3.uvm.edu. 2022. *Gamma Distribution Explained | What is Gamma Distribution?*. [online] Available at: <https://pdodds.w3.uvm.edu/files/papers/others/1986/hastie1986a.pdf> [Accessed 16 September 2022].

project, T., 2022. *Contacting the World Air Quality Index team*. [online] aqicn.org. Available at: <https://aqicn.org/contact/#:~:text=Its%20mission%20is%20to%20promote,.org%20and%20waqi.info.> [Accessed 16 September 2022].

Publichealth.columbia.edu. 2022. Geographically Weighted Regression | Columbia Public Health. [online] Available at: <https://www.publichealth.columbia.edu/research/population-health-methods/geographically-weighted-regression#:~:text=Geographically%20weighted%20regression%20(GWR)%20is,and%20an%20outcome%20of%20interest.> [Accessed 16 September 2022].

QuestionPro. 2022. *Empirical Research: Definition, Methods, Types and Examples | QuestionPro*. [online] Available at: <https://www.questionpro.com/blog/empirical-research/#Types_and_methodologies_of_empirical_research> [Accessed 16 September 2022].

Sagepub.com. 2022. *CHAPTER 2. SPECIFICATION OF SIMULTANEOUS EQUATION MODELS*. [online] Available at: <https://www.sagepub.com/sites/default/files/upm-binaries/39916_Chapter2.pdf> [Accessed 16 September 2022].

Scholar.pku.edu.cn. 2022. *IMED Overview*. [online] Available at: <http://scholar.pku.edu.cn/hanchengdai/imed_general#:~:text=IMED%20is%20a%20system%20of,provide%20relevant%20scientific%20support%20for> [Accessed 16 September 2022].

Soleimani, M., Akbari, N., Saffari, B. and Haghshenas, H., 2022. Health effect assessment of PM2.5 pollution due to vehicular traffic (case study: Isfahan). *Journal of Transport &amp; Health*, 24, p.101329.

Statistics How To. 2022. *ARMA model*. [online] Available at: <https://www.statisticshowto.com/arma-model/> [Accessed 16 September 2022].

Statistics How To. 2022. *Simultaneous Equations Model (SEM): Simple Definition*. [online] Available at: <https://www.statisticshowto.com/simultaneous-equations-model/> [Accessed 16 September 2022].

Statistics Solutions. 2022. *What is Linear Regression? - Statistics Solutions*. [online] Available at: <https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/what-is-linear-regression/> [Accessed 16 September 2022].

Swaraj, A., Verma, K., Kaur, A., Singh, G., Kumar, A. and Melo de Sales, L., 2021. Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India. *Journal of Biomedical Informatics*, 121, p.103887.

Tableau. 2022. Time Series Analysis: Definition, Types, Techniques, and When It's Used. [online] Available at: <https://www.tableau.com/learn/articles/time-series-analysis> [Accessed 11 September 2022].

The World Bank Group. 2016. China - Innovative Financing for Air Pollution Control in Jing-Jin-Ji Project. [online] Available at: <https://documents.worldbank.org/curated/en/488161468187136819/pdf/102272-PAD-P154669-R2016-0031-1-OUO-9.pdf> [Accessed 11 September 2022].

The World Bank. 2019. *China's Hebei Province Fights for Blue Skies with World Bank Support*. [online] Available at: <https://www.worldbank.org/en/news/feature/2019/06/05/chinas-hebei-province-fights-for-blue-skies-with-world-bank-support#:~:text=The%20province%20has%20the%20highest,with%20a%20large%20agricultural%20sector.> [Accessed 16 September 2022].

The World Bank. 2020. China: Fighting Air Pollution and Climate Change through Clean Energy Financing. [online] Available at: <https://www.worldbank.org/en/results/2020/06/21/china-fighting-air-pollution-and-climate-change-through-clean-energy-financing> [Accessed 6 September 2022].

Vle.exeter.ac.uk. 2022. *MTHM506 Statistical Data Modelling Topic 3- Another example of GAMs*. [online] Available at: <https://vle.exeter.ac.uk/mod/resource/view.php?id=2204810> [Accessed 16 September 2022].

Wang, G., Gu, S., Chen, J., Wu, X. and Yu, J., 2016. Assessment of health and economic effects by PM2.5pollution in Beijing: a combined exposure–response and computable general equilibrium analysis. *Environmental Technology*, 37(24), pp.3131-3138.

Wang, J., Zhang, L., Niu, X. and Liu, Z., 2020. Effects of PM2.5 on health and economic loss: Evidence from Beijing-Tianjin-Hebei region of China. *Journal of Cleaner Production*, 257, p.120605.

Web.pdx.edu. 2022. *Applied Spatial Econometrics: Raising the Bar*. [online] Available at: <https://web.pdx.edu/~crkl/SEAUG/papers/Elhorst_SEA2010.pdf> [Accessed 16 September 2022].

Ww2.arb.ca.gov. 2022. Inhalable Particulate Matter and Health (PM2.5 and PM10) | California Air Resources Board. [online] Available at: <https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health#:~:text=It%20is%20a%20complex%20mixture,solid%20cores%20with%20liquid%20coatings.> [Accessed 11 September 2022].

Xiao, Q., Geng, G., Liang, F., Wang, X., Lv, Z., Lei, Y., Huang, X., Zhang, Q., Liu, Y. and He, K., 2020. Changes in spatial patterns of PM2.5 pollution in China 2000–2018: Impact of clean air policies. Environment International, [online] 141, p.105776. Available at: <https://www.sciencedirect.com/science/article/pii/S0160412020309302>.

Xie, Y., Dai, H., Dong, H., Hanaoka, T. and Masui, T., 2016. Economic Impacts from PM2.5 Pollution-Related Health Effects in China: A Provincial-Level Analysis. Environmental Science &amp; Technology, [online] 50(9), pp.4836-4843. Available at: <https://pubs.acs.org/doi/full/10.1021/acs.est.5b05576>.

Xie, Y., Dai, H., Dong, H., Hanaoka, T. and Masui, T., 2016. Economic Impacts from PM2.5 Pollution-Related Health Effects in China: A Provincial-Level Analysis. Environmental Science &amp; Technology, [online] 50(9), pp.4836-4843.Available at: <https://www.researchgate.net/publication/301217550_Economic_Impacts_from_PM25_Pollution-Related_Health_Effects_in_China_A_Provincial-Level_Analysis>.

Xie, Y., Dai, H., Zhang, Y., Wu, Y., Hanaoka, T. and Masui, T., 2019. Comparison of health and economic impacts of PM2.5 and ozone pollution in China. *Environment International*, 130, p.104881.

Xu, X. and Zhang, T., 2020. Spatial-temporal variability of PM2.5 air quality in Beijing, China during 2013–2018. *Journal of Environmental Management*, 262, p.110263.

Yan, D., Kong, Y., Jiang, P., Huang, R. and Ye, B., 2021. How do socioeconomic factors influence urban PM2.5 pollution in China? Empirical analysis from the perspective of spatiotemporal disequilibrium. Science of The Total Environment, [online] 761, p.143266. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0048969720367978#bb0050>.

Zhang, L., Lin, J., Qiu, R., Hu, X., Zhang, H., Chen, Q., Tan, H., Lin, D. and Wang, J., 2018. Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model. *Ecological Indicators*, 95, pp.702-710.

Appendix: Code

Packages

```
# Loading the required packages
library(ggplot2)
library(scales)        # to adjust the date in the graphs
library(zoo)
library(dplyr)          # to convert daily data to quarterly
library(mgcv)           # to fit GAM model
library(MASS)          # to fit negative binomial GAM model
library(forecast)       # to fit ARIMA model
```

PM2.5 Pollution Dataset

```
#### Importing and wrangling the data ####
#Loading the pm2.5 pollution Dataset

> china_pm2.5 <- read.csv("C:/Users/91988/OneDrive/Desktop/Project/PM2.5 data/PM2.5 Pollution
data.csv")
#Loading the location data
>  loc  <-  read.csv("C:/Users/91988/OneDrive/Desktop/Project/PM2.5  data/aqi_locations_2016-02-
04.csv")

#Changing the date from character format to date format
> china_pm2.5$date <- as.Date(china_pm2.5$date, "%d-%m-%Y")

#Changing province column format from character to Factor
> china_pm2.5$Province <- as.factor(china_pm2.5$Province)

#Selecting only pm2.5 column
> china_pm2.5 <- china_pm2.5[,c(1,2,3,4)]

#Selecting the time period between 2014 to 2021
> china_pm2.5 <- china_pm2.5[china_pm2.5$date > "2013-12-31" &
                        china_pm2.5$date < "2022-01-01",]

#merging the two files to get the coordinates
> pm2.5_data <- merge(china_pm2.5, loc[,1:4], by.x = "AQI.Locations",
                by.y = "stationname")

#Converting the daily data to Quarterly
#Separating the year from the date column
> pm2.5_data["Year"] <- format(pm2.5_data$date, format = "%Y")
> pm2.5_data$Quarter <- as.yearqtr(pm2.5_data$date)

#Finding the quantiles for the provinces (quarterly)
> pm2.5_quant <- pm2.5_data %>%
  group_by(Province, Quarter) %>%
  summarize(count = n(),
          min = fivenum(pm25)[1],
          median = fivenum(pm25)[2],
          max = fivenum(pm25)[3],
          sum = sum(pm25, na.rm = TRUE),
          mean = mean(pm25, na.rm = TRUE))

> pm25_qrtly <- pm2.5_data %>%
  group_by(Quarter, Province, Year) %>%
  summarize(PM25 = mean(pm25, na.rm = TRUE))
```

Initial Data Analysis (PM2.5)

For PM2.5 (daily levels)

```
#numerical analysis
> summary(pm2.5_data)

#graphical analysis
> ggplot(data = pm2.5_data, aes(x = date, y = pm25, color = factor(Province)))+
  geom_line()+
  labs(x = 'Year',y = 'PM2.5',
     title = 'PM2.5 Pollution (2014-2021)', color = 'Province/Municipality')+
  facet_wrap(.~ Province, ncol = 3, scales = "free")
```

For PM2.5 (Quarterly levels)

```
#numerical analysis
> summary(pm25_qrtly)

#Graphical analysis
> ggplot(data = pm2.5_data, aes(x = Quarter, y = pm25, color = factor(Province)))+
  geom_point()+
  labs(x = 'Quarter',y = 'PM2.5',
     title = 'PM2.5 Pollution (2014-2021)', color = 'Province/Municipality')+
  facet_wrap(.~ Province, ncol = 3, scales = "free")
```

GDP Dataset

```
#Loading the economic variables dataset
> gdp_data <- read.csv("C:/Users/91988/OneDrive/Desktop/Project/Economy Variables/GDP
Quarterly.csv")

#Changing City column format from character to Factor
> gdp_data$Province <- as.factor(gdp_data$Province)

#Changing the Year from integer format to date format
> gdp_data$Year <- as.Date(as.character(gdp_data$Year), format = "%Y")

#Separating the year from the date column
> gdp_data$Year <- format(gdp_data$Year, format = "%Y")

#removing the comma's from the columns in order to convert
```

```
#the class from character to numeric
> gdp_data$GDP..100.million.yuan. <- as.numeric(gsub(",","",gdp_data$GDP..100.million.yuan.))

#Changing column names
> gdp_data <- gdp_data %>% rename(GDP = GDP..100.million.yuan.)

#Finding the quantiles for the provinces (quarterly)
> GDP_quant <- gdp_data %>%
  group_by(Province) %>%
  summarize(count = n(),
          min = fivenum(GDP)[1],
          median = fivenum(GDP)[2],
          max = fivenum(GDP)[3],
          sum = sum(GDP, na.rm = TRUE),
          mean = mean(GDP))
```

Initial Data Analysis (GDP)

```
#summary of the gdp data
> summary(gdp_data)

#Graphical analysis of GDP
> ggplot(data = gdp_data, aes(x = Quarter, y = GDP)) +
  geom_point(aes(colour = Province), size = 1)+
  labs(color ='Province/Municipality', title = "Quarterly GDP from 2014-2021") +
  ylab("GDP") + xlab("Quarter") +
  theme(axis.text.x = element_text(angle = 90))
```

Regression Analysis (GAM Model)

```
#Combining the pm2.5 data with the gdp data
#Changing the yearqtr format to character in the pm2.5 data
> pm25_qrtly$Quarter <- as.character(pm25_qrtly$Quarter)
> china_data <- pm25_qrtly %>% full_join(gdp_data)

#Graphical summary
> pairs(~GDP + PM25, data = china_data)

#Plotting the data
> plot(china_data$PM25, china_data$GDP, pch=20,xlab='PM2.5',ylab='GDP',
    main = 'GDP vs PM25 from 2014-2021')

#Histogram for PM2.5
> hist(china_data$PM25)

#Histogram for GDP
> hist(china_data$GDP)

#Fitting a LM model
> lm_mod <- lm(GDP ~ PM25, data = china_data)
> summary(lm_mod)
#Scatterplot with regression line
> ggplot(china_data, aes(y=GDP, x=PM25)) +geom_point()+
  geom_smooth(method = 'lm')+
  labs(title = "Linear Regression Scatter Plot") +
  ylab("PM2.5") + xlab("GDP")

#Fitting a GAM Gaussian model
> gam_mod <- gam(GDP ~ s(PM25, by = Province, k=30, bs='cs'),
              data = china_data,
              method = "REML",
              family = gaussian(link = "identity"))
> summary(gam_mod)

#Checking residuals
> par("mar")
> par(mar=c(2,2,2,2))
> par(mfrow = c(2,2))
> gam.check(gam_mod,pch=20)

#Fitting a poisson GAM model
> gam_mod2 <- gam(GDP ~ s(PM25, by = Province, k=30, bs='cs'),
              data = china_data,
              family = poisson(link = "log"))
> summary(gam_mod2)

#Checking residuals
> par(mar=c(2,2,2,2))
> par(mfrow = c(2,2))
> gam.check(gam_mod2,pch=20)

#Fitting a negative binomial GAM model
> gam_mod3 <- gam(GDP ~ s(PM25, by = Province, k=30, bs='cs'),
               data = china_data,
               method = "REML",
               family = negbin(352))
> summary(gam_mod3)

#Checking residuals
> par(mar=c(2,2,2,2))
> par(mfrow = c(2,2))
> gam.check(gam_mod3,pch=20)

#Fitting a quasipoisson GAM model
> gam_mod4 <- gam(GDP ~ s(PM25, by = Province, k=30, bs='cs'),
               data = china_data,
               method = "REML",
               family = quasipoisson(link = "log"))
> summary(gam_mod4)
```

```
#Checking residuals
> par(mar=c(2,2,2,2))
> par(mfrow = c(2,2))
> gam.check(gam_mod4,pch=20)

#Fitting a gamma GAM model
> gam_mod5 <- gam(GDP ~ s(PM25, by = Province, k=30, bs='cs'),
            data = china_data,
            method = "REML",
            family = Gamma(link = "log"))
> summary(gam_mod5)

#Checking residuals
> par(mar=c(2,2,2,2))
> par(mfrow = c(2,2))
> gam.check(gam_mod5,pch=20)

#Extracting the AIC's
> AIC(gam_mod, gam_mod2, gam_mod3, gam_mod4, gam_mod5)
```

Time Series Analysis (ARIMA models)

```
##TIME SERIES FOR PM2.5 for the provinces
#creating time series data for PM2.5
> ts_data <- china_data[,c("Quarter","PM25")]

#graphical summary
> ggplot(data = ts_data, aes(x = Quarter, y = PM25)) +
  geom_point()+
  labs(title = "PM2.5 Concentration from 2014-2021") +
  ylab("PM2.5") + xlab("Quarter") +
  theme(axis.text.x = element_text(angle = 90))

#numerical summary
> summary(ts_data)

#Periodogram
> spec.pgram(ts_data, log = 'no', na.action = na.pass)

#converting the PM2.5 data to a TS data
> qrtly.ts <- ts(ts_data[,'PM25'], start=c(2014,1), end = c(2021,4), frequency = 4)
> qrtly.ts

#plotting the ts data
> plot(qrtly.ts, lwd=2, col = 'blue', xlab='Year', ylab = 'Quarterly PM2.5 Concentration',
    main = 'PM2.5 Time Series Plot')

#ARIMA
```

```
> par(mfrow=c(1,2))
> plot(qrtly.ts)    #before differencing
> plot(diff(qrtly.ts))   #after differencing

#Checking the ACF and PACF
> par(mfrow=c(1,2))
> acf(diff(qrtly.ts), main = 'ACF'); pacf(diff(qrtly.ts), main = 'PACF')

#Automatically estimating the orders
> auto.arima(qrtly.ts, seasonal = FALSE)

#ARIMA Model 1 – AR(1,1,0)
> arima1 <- arima(qrtly.ts, order = c(1,1,0))
> arima1

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(arima1)

#ARIMA Model 2 – MA(0,1,1)
> arima2 <- arima(qrtly.ts, order = c(0,1,1))
> arima2

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(arima2)

#ARIMA Model 3 – ARIMA(1,1,1)
> arima3 <- arima(qrtly.ts, order = c(1,1,1))
> arima3

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(arima3)

#ARIMA Model 4 – ARIMA(2,1,0)
> arima4 <- arima(qrtly.ts, order = c(2,1,0))
> arima4

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(arima4)

#ARIMA Model 5 – ARIMA(0,1,2)
> arima5 <- arima(qrtly.ts, order = c(0,1,2))
> arima5

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(arima5)
### PREDICTIONS ###
#PM2.5
> pred_PM2.5 <- predict(arima3, n.ahead = 50)
> pred_PM2.5$pred

#Plotting the predictions
> plot(qrtly.ts, xlab = "Quarter", ylab = "PM2.5 Concentration",
     main = "PM2.5 Prediction for Chinese Provinces")
> lines(pred_PM2.5$pred, col = "blue", lwd=2)
> lines(pred_PM2.5$pred+1.96*pred_PM2.5$se, col = "red", lwd=2)
> lines(pred_PM2.5$pred-1.96*pred_PM2.5$se, col = "red", lwd=2)


##TIME SERIES FOR GDP of the Provinces

#Creating GDP time series data
> gdpts_data <- china_data[,c("Quarter","GDP")]

#Graphical summary
> ggplot(data = gdpts_data, aes(x = Quarter, y = GDP)) +
  geom_point()+
  labs(title = "PM2.5 Concentration from 2014-2021") +
  ylab("GDP") + xlab("Quarter") +
  theme(axis.text.x = element_text(angle = 90))

#Numerical sumary
> summary(gdpts_data)

#Periodogram
> spec.pgram(gdpts_data, log = 'no')

#converting the PM2.5 data to a TS data
> gdp_qrtly.ts <- ts(gdpts_data[,'GDP'], start=c(2014,1), end = c(2021,4), frequency = 4)
> gdp_qrtly.ts

#plotting the ts data
> plot(gdp_qrtly.ts, lwd=2, col = 'blue', xlab='Quarter', ylab = 'Quarterly GDP',
     main = 'GDP Time Series Plot')

#ARIMA
> par(mfrow=c(1,2))
> plot(gdp_qrtly.ts)    #before differencing
> plot(diff(gdp_qrtly.ts))   #after differencing

#Checking ACF & PACF
> par(mfrow=c(1,2))
> acf(diff(gdp_qrtly.ts), main = 'ACF'); pacf(diff(gdp_qrtly.ts), main = 'PACF')

#Automatically estimating the orders
```

```
> auto.arima(gdp_qrtly.ts, seasonal = FALSE)
#ARIMA Model 1 – AR(1,1,0)
> gdp_arima1 <- arima(gdp_qrtly.ts, order = c(1,1,0))
> gdp_arima1

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(gdp_arima1)

#ARIMA Model 2 – MA(0,1,1)
> gdp_arima2 <- arima(gdp_qrtly.ts, order = c(0,1,1))
> gdp_arima2

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(gdp_arima2)

#ARIMA Model 3 – ARIMA(1,1,1)
> gdp_arima3 <- arima(gdp_qrtly.ts, order = c(1,1,1))
> gdp_arima3

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(gdp_arima3)

#ARIMA Model 4 – ARIMA(2,1,0)
> gdp_arima4 <- arima(gdp_qrtly.ts, order = c(2,1,0))
> gdp_arima4

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(gdp_arima4)

#ARIMA Model 5 – ARIMA(0,1,2)
> gdp_arima5 <- arima(gdp_qrtly.ts, order = c(3,1,0))
> gdp_arima5

#Residual diagnostics
> par(mar=c(2,2,1,1))
> tsdiag(gdp_arima5)

## PREDICTIONS ##
#GDP
> pred_gdp <- predict(gdp_arima4, n.ahead = 50)
> pred_gdp$pred

#Plotting the predictions
> plot(gdp_qrtly.ts, xlab = "Quarter", ylab = "GDP",
    main = "GDP Prediction for Chinese Provinces")
```

```
> lines(pred_gdp$pred, col = "blue", lwd=2)
> lines(pred_gdp$pred+1.96*pred_gdp$se, col = "red", lwd=2)
> lines(pred_gdp$pred-1.96*pred_gdp$se, col = "red", lwd=2)
```

Additional important Tables & Figures

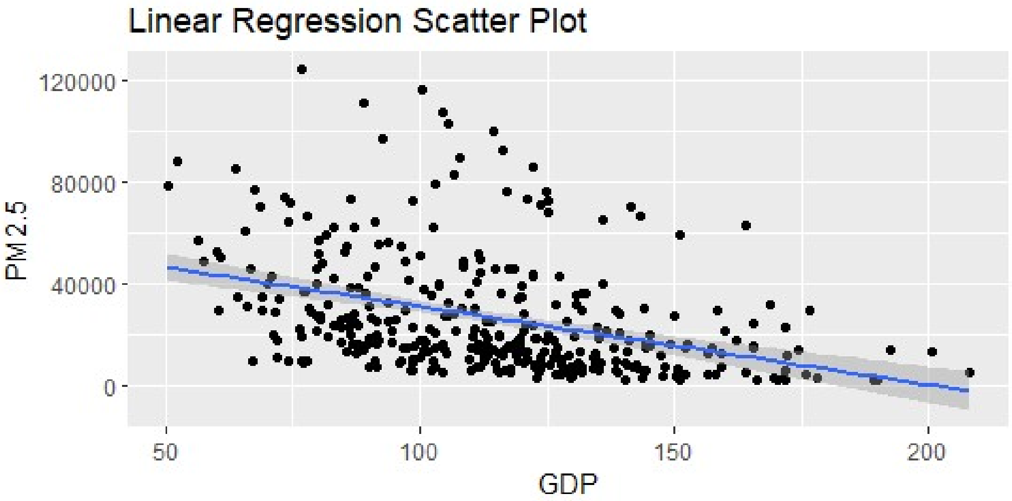

Figure 6: Linear Regression Scatter plot

Table 6: Results produced by gam.check() for Gamma GAM model

	k'	*edf*	*k-index*	*p-value*
s(PM25): Province Beijing	29	1.0328	0.94	0.29
s(PM25): Province Chongqing	29	1.1255	0.94	0.3
s(PM25): Province Guangdong	29	3.5672	0.94	0.28
s(PM25): Province Hebei	29	0.0137	0.94	0.26
s(PM25): Province Hubei	29	2.331	0.94	0.32
s(PM25): Province Jiangsu	29	9.2706	0.94	0.34
s(PM25): Province Jilin	29	11.2993	0.94	0.32
s(PM25): Province Shandong	29	12.8636	0.94	0.26
s(PM25): Province Shanghai	29	1.0282	0.94	0.3
s(PM25): Province Tianjin	29	2.0282	0.94	0.34
s(PM25): Province Zhejiang	29	2.3437	0.94	0.37

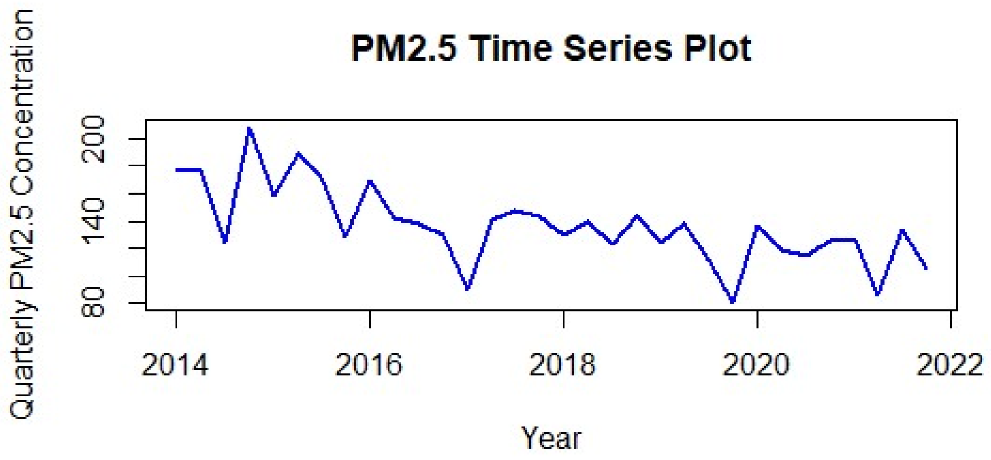

Figure 7: PM2.5 quarterly time series plot

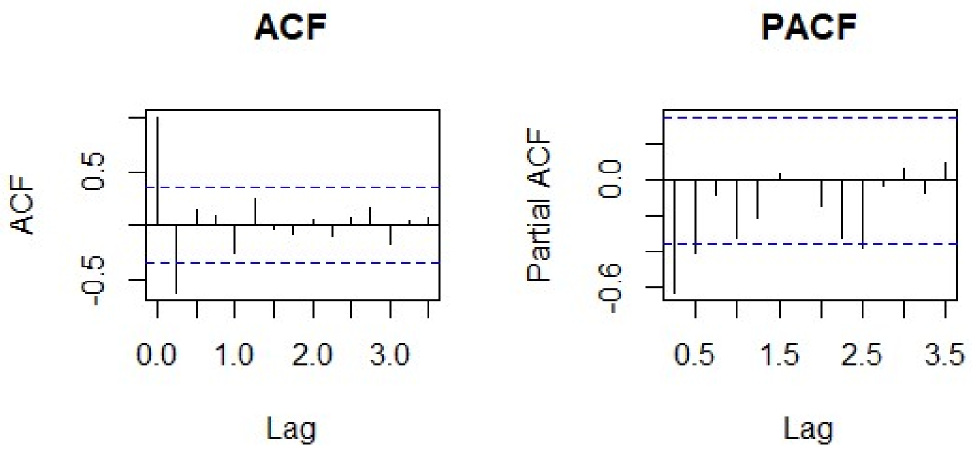

Figure 8: PM2.5 ACF & PACF before and after differencing

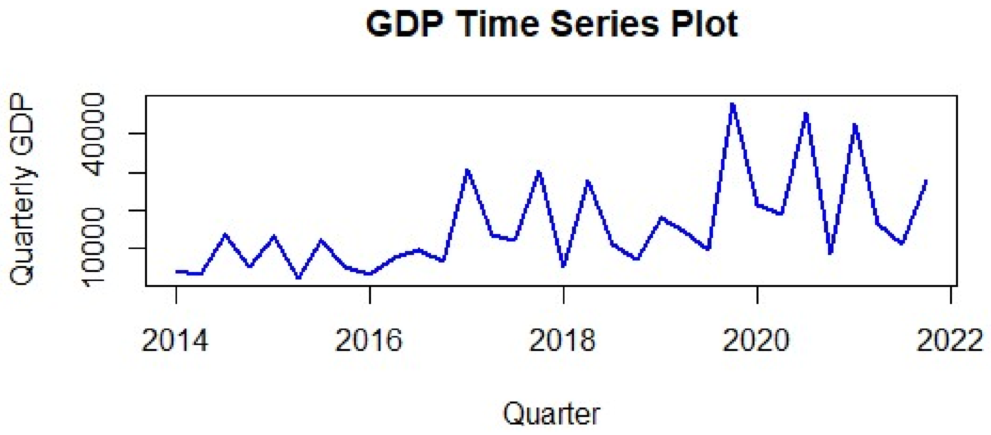

Figure 9: GDP quarterly time series plot

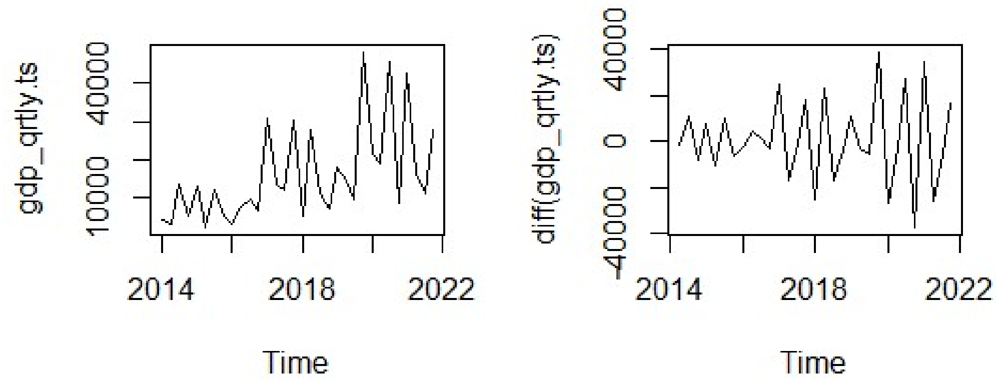

Figure 10: GDP ACF & PACF before and after differencing

Table 7: Predicted levels of PM2.5 till 2Q of 2034

	Qtr1	*Qtr2*	*Qtr3*	*Qtr4*
2022	117.7193	113.8316	115.0788	114.6787
2023	114.807	114.7659	114.7791	114.7748
2024	114.7762	114.7758	114.7759	114.7758
2025	114.7759	114.7759	114.7759	114.7759
2026	114.7759	114.7759	114.7759	114.7759
2027	114.7759	114.7759	114.7759	114.7759
2028	114.7759	114.7759	114.7759	114.7759
2029	114.7759	114.7759	114.7759	114.7759
2030	114.7759	114.7759	114.7759	114.7759
2031	114.7759	114.7759	114.7759	114.7759
2032	114.7759	114.7759	114.7759	114.7759
2033	114.7759	114.7759	114.7759	114.7759
2034	114.7759	114.7759		

Table 8: Predicted values of GDP till 2Q of 2034

	Qtr1	*Qtr2*	*Qtr3*	*Qtr4*
2022	12379.93	19017.6	21572.97	14312.86
2023	20909.83	18150.8	16971.88	20122.64
2024	17305.49	18450.66	18991.74	17625.01
2025	18827.58	18352.99	18105.78	18698.36
2026	18185.22	18381.57	18494.07	18237.26
2027	18456.13	18375.04	18324.02	18435.27
2028	18341.95	18375.38	18398.44	18350.27
2029	18390.04	18376.29	18365.9	18386.75
2030	18369.8	18375.45	18380.12	18371.1

2031	18378.31	18376	18373.91	18377.81
2032	18374.74	18375.68	18376.62	18374.93
2033	18376.24	18375.85	18375.44	18376.16
2034	18375.61	18375.76		

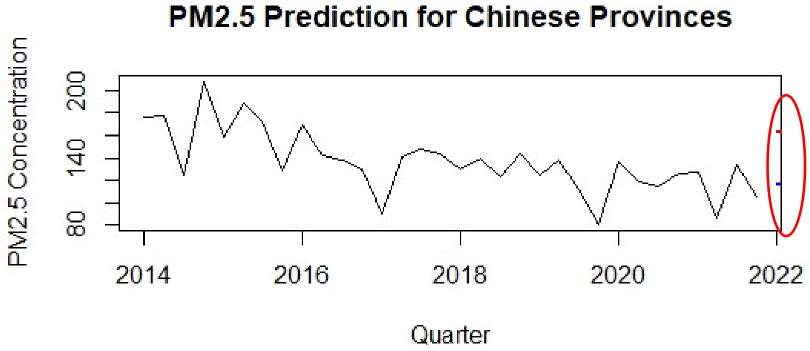

Figure 11: PM2.5 time series prediction plot

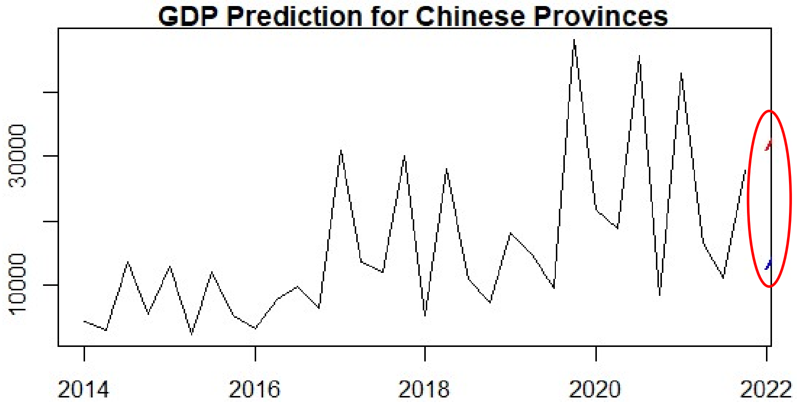

Figure 12: GDP time series prediction plot