SAINT-PETERSBURG STATE UNIVERSITY

Qi Dongfang

Research on Investment Attractiveness and Environmental Safety in China and Southeast Asia: Empirical Models and Data Analysis

Scientific specialty – 1.2.2.Mathematical modeling, numerical methods and software packages.

DISSERTATION

Thesis for the degree of Candidate of Technical Sciences

Scientific advisor:

Doctor of Technical Sciences, professor

Bure Vladimir Mansurovich

Saint-Petersburg

2024

Contents

Introduction

The relevance of the thesis topic

The research topic explored in this thesis holds significant relevance from both academic and practical perspectives. By investigating the relationship between various statistical models and research in the fields of economics and ecology, specifically focusing on China and Southeast Asia, this study addresses crucial issues that have implications for economic development and environmental sustainability in the region. Analyzing the investment attractiveness of China and ASEAN5 using multiple linear regression is of great importance given the rapid growth and increasing importance of these economies. Understanding the factors that influence investment attractiveness provides valuable insights for policymakers, businesses, and investors seeking to allocate resources effectively. The utilization of stepwise regression further enhances the analysis by identifying the most important factors among a wide range of potential variables, helping stakeholders prioritize their decision-making processes.

Segmenting China into four distinct groups based on investment attractiveness allows for more nuanced analyses and assessments of regional differences. Understanding the unique characteristics and challenges faced by each group can inform tailored strategies for economic development and environmental conservation. Policymakers can utilize the findings to design targeted policies that account for regional disparities and optimize resource allocation. Analyzing the determinants of air quality through stepwise regression addresses the pressing ecological concern of air pollution. Identifying key contributing factors facilitates evidence-based decision-making in environmental management.

Furthermore, applying advanced neural network models such as ANN, RNN, LSTM, GRU, BiRNN, BiLSTM, and BiGRU, along with ensemble models like XGBoost, LightGBM, and CatBoost, improves air quality prediction accuracy. These techniques capture complex temporal patterns and provide a foundation for developing forecasting systems that assist stakeholders in making informed decisions to address air pollution challenges. Utilizing Explainable AI techniques, such as SHAP values, enhances transparency and interpretability of black box models. Understanding the factors contributing to model predictions is critical for building trust, gaining insights, and making informed decisions based on the models' outputs. This analysis aligns with the growing demand for explainable and accountable AI systems, addressing concerns raised by stakeholders regarding the use of complex machine learning models in real-world applications.

In conclusion, the research topic addressed in this thesis holds significant relevance to both academia and practice. By investigating the relationship between statistical models and research in economics and ecology in China and Southeast Asia, the study provides valuable insights into investment attractiveness, air quality analysis, regional disparities, time series prediction, ensemble modeling, and explainability of black box models. The findings contribute to knowledge advancement, offer practical implications for decision-makers, and facilitate evidence-based policy formulation in the fields of economics and ecology.

Overview of achievements in the field

The research conducted in this thesis builds upon the significant contributions made by scholars and researchers in the fields of investment analysis and environmental science. The advancements in understanding investment attractiveness, air quality analysis, regional disparities, time series prediction, model selection, and interpretability of complex models have provided a foundation for this study.

Investment attractiveness holds a crucial role in investor decision-making processes, facilitating risk mitigation, return maximization, and efficient capital allocation. By evaluating various factors such as market conditions, competition, regulations, and economic stability, investors can make informed decisions aligned with their financial goals. Modeling investment attractiveness enhances evaluation by offering a comprehensive framework. These models incorporate both quantitative and qualitative variables, enabling forecasting, scenario analysis, and comparative assessment. These approaches reduce reliance on subjective judgments and improve the accuracy of investment decisions. Acknowledging the significance of investment attractiveness and utilizing modeling techniques contributes to optimizing portfolio performance and fostering sustainable economic growth.

Linear regression and stepwise regression analysis are widely used methods for modeling investment attractiveness. These approaches provide simplicity, interpretability, and statistical significance testing, enabling a systematic analysis of factors influencing investment attractiveness. Linear regression allows for the inclusion of quantitative data and facilitates the identification of relationships between independent and dependent variables. Furthermore, hypothesis testing provides insights into the statistical significance of observed associations, reinforcing robust conclusions. Stepwise regression analysis extends linear regression by automatically selecting relevant independent variables, enhancing interpretability, and computational efficiency.

In the second chapter, cluster analysis is employed to identify distinct groups or clusters based on similarities and dissimilarities among selected variables. This analysis reveals underlying patterns and structures within the dataset, providing valuable insights into the relationships and characteristics of the investment entities examined.

The article [1] aims to explore the methodological support for assessing the investment attractiveness of innovative companies, addressing the information needs of stakeholders. The authors employ analysis and synthesis methods to define and structure the concept of investment attractiveness and its analytical characteristics. Paper [2] aims to identify and validate the theoretical characteristics of an investment-attractive city by analyzing economic literature and surveying entrepreneurs. The key factors found to influence the choice of a city for investment include accessibility of skilled workforce, labor costs, resource prices, and market competition. These articles [3, 4, 5, 6] explore diverse aspects of investment attractiveness, ranging from analyzing specific countries' performance and rankings to examining investment strategies and addressing the challenges and perspectives for enhancing investment attractiveness.

Scholars using statistical methods to analyze investment attractiveness may encounter limitations, including oversimplification, a lack of consideration for contextual factors, and inherent assumptions. These limitations can lead to inaccurate predictions and an incomplete understanding of the drivers of investment attractiveness.

However, stepwise regression analysis offers advantages and innovations in addressing these limitations. By automatically selecting relevant variables, it overcomes oversimplification and allows for the identification of non-linear relationships and contextual factors. The flexibility of the model enables iterative refinement, adapting to changing market conditions and incorporating emerging trends. Through these advancements, stepwise regression analysis provides novel insights into the factors driving investment attractiveness, contributing to a deeper understanding of decision-making. This approach aligns with academic rigor, logical reasoning, and the standards of scholarly research, enhancing prediction accuracy and providing valuable insights into the complexities of investment attractiveness.

The field of environmental science has made significant progress in analyzing air quality and its implications for public health and the environment. Studies have been conducted to investigate the spatiotemporal patterns of air pollution in China using Air Quality Index (AQI) data, revealing high pollution levels across the country and identifying PM2.5, PM10, and O3 as major pollutants [7]. Forecasting air quality, particularly pollution parameters, has become crucial for decision-making processes in this domain.

Researchers have developed models to predict daily AQI, adapting methods similar to those used by the US Environmental Protection Agency (USEPA) to Indian standards [8]. [9] finds no significant variation in AQI during weekends versus weekdays, treating all days equally in the models. Additionally, the interconnectedness of air pollution and climate change has been emphasized, calling for coordinated actions that address their linkages [10].

Studies have identified the predominance of natural factors over socioeconomic factors in influencing air pollution, with interactions among driving factors leading to nonlinear-enhanced or bi-enhanced effects [11, 12, 13, 14, 15]. These findings have significant policy implications for mitigating air pollution in China.

Previous research has explored various air pollutants, such as particulate matter, ozone, carbon monoxide, sulfur dioxide, and nitrogen dioxide, to understand their sources, dispersion patterns, and health effects [16, 17, 18, 19, 20]. Statistical techniques, including regression analysis, time series forecasting, and machine learning algorithms, have been applied to model and predict air quality. These studies([21, 22, 23, 24]) have deepened our understanding of regional disparities, identified key factors impacting development, and offered practical implications for decision-makers in the field.

These advancements in understanding air pollution patterns and forecasting contribute to improved management strategies and policy development. They shed light on the complex interactions among driving factors and provide insights into the sources, dispersion, and health effects of air pollutants, enabling more effective mitigation efforts.

In recent years, there have been significant advancements in time series prediction using artificial neural networks (ANNs). Researchers have explored various architectures such as recurrent neural networks (RNN), long short-term memory (LSTM), gated recurrent units (GRU), and their variants to model and predict time-dependent data. These models have demonstrated their effectiveness in capturing complex temporal patterns and forecasting future values in paper [25, 26, 27, 28]. For improved accuracy in time series forecasting, a hybrid methodology proposed by Zhang et al. [29] combines autoregressive integrated moving average (ARIMA) and ANN models.

The application of ANNs has also been successful in environmental science. Palani et al. [30] demonstrate the use of ANN models in predicting water quality variables in Singapore coastal waters, accurately simulating salinity, temperature, dissolved oxygen, and chlorophyll-a levels.

Research [31, 32, 33] provides comprehensive tutorials on RNNs and LSTMs, explaining the derivation of equations, addressing training difficulties, and introducing enhanced versions of LSTM models. These resources offer valuable insights for researchers looking to implement augmented LSTM models.

A review by De Gooijer and Hyndman [34] covers 25 years of research in time series forecasting, focusing on studies published in journals managed by the International Institute of Forecasters. The review highlights significant contributions, identifies areas for further development, and suggests future research directions.

Lim et al. [35] provide an extensive overview of time series forecasting, emphasizing the increasing use of deep neural networks. They discuss common deep learning architectures, including feed forward networks, recurrent neural networks (Elman, LSTM, GRU, bidirectional), and convolutional neural networks. Practical aspects such as hyper-parameter settings and framework choices are also explored.

In the context of financial time series forecasting, Sezer et al. [36] present a comprehensive review of studies utilizing deep learning models. Categorizing implementations by domain (index, forex, commodity) and deep learning model choices (CNNs, DBNs, LSTM), this review provides insights into the potential and limitations of deep learning models in financial forecasting.

The application of ANNs extends to pharmaceutical sciences, as highlighted by Agatonovic-Kustrin et al. [37]. ANNs simulate the information processing of the human brain, enabling applications in classification, prediction, and modeling, supporting drug design and clinical pharmacy.

Furthermore, Lim et al. [38] survey deep learning architectures in time-series forecasting, discussing encoder and decoder designs for one-step-ahead and multi-horizon forecasting. The integration of statistical models with neural networks in hybrid models is explored, along with the potential benefits of deep learning in decision-making with time-series data. These findings contribute to the understanding and application of deep learning techniques in forecasting time-dependent data.

These papers [39, 40, 41] demonstrate the strides made in time series forecasting through the application of boosting models like XGBoost, LightGBM, and CatBoost. These models have demonstrated their superiority over traditional statistical methods in domains such as air quality prediction. They effectively capture complex relationships, nonlinear patterns, and handle categorical features. The incorporation of these advanced boosting models enhances the accuracy and reliability of air quality predictions, supporting informed decision-making by policymakers and stakeholders.

Research by Sagi et al. [42] addresses the need for interpretable machine learning models and proposes a method to transform GBDT models into interpretable decision trees without sacrificing predictive performance. A study by Ramraj et al. [43] compares the accuracy and speed of XGBoost with traditional Gradient Boosting in multi-threaded single-system mode, demonstrating the superior training time and performance of XGBoost. The work [44, 45, 46] introduces XGBoost, a scalable tree boosting system widely used in machine learning tasks, with novel algorithms for handling sparse data and approximate tree learning. Similarly, the article [47] proposes Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques to improve efficiency and scalability in Gradient Boosting Decision

Tree (GBDT) algorithms, resulting in the development of LightGBM.

Other relevant articles, such as those on CatBoost and its applications [48, 49, 50, 51], as well as studies on decision tree algorithms [52, 53], contribute to the broader understanding of boosting models in machine learning. Paper [54] focuses on predicting hourly PM2.5 concentration in China using the XGBoost algorithm. The study evaluates the performance of XGBoost by comparing observed and predicted PM2.5 concentrations, demonstrating its superiority over other data mining methods. In the article by Ju et al. [55], a model combining convolutional neural network (CNN) and LightGBM algorithm is proposed for ultra-short-term wind power forecasting. This approach leverages the strengths of both models to achieve improved accuracy in wind power prediction.

The article by Dorogush et al. [56] introduces CatBoost, a gradient boosting framework with support for categorical features. It highlights the advantages of CatBoost in handling datasets with categorical features, offering improved performance compared to traditional gradient boosting models.

These studies further contribute to the field of time series prediction by exploring the application of XGBoost, LightGBM, and CatBoost algorithms in different domains, including air quality forecasting and wind power prediction, showcasing their effectiveness and superiority over other methods.

In terms of explainable AI, SHAP value analysis has made significant contributions to model interpretability. Research studies by Marcilio et al. [57], Meng et al. [58], and Mokhtari et al. [59] explore the application of SHAP values as feature selection mechanisms and interpretable solutions across various domains.

These research studies highlight the significance of SHAP values in model interpretation across various domains, including process management in wastewater treatment plants [60] and machine learning explanations [61, 62, 63]. The body of literature surrounding SHAP values includes seminal works such as those by Winter [64] and Roth [65], as well as studies that explore variations and algorithmic approaches [66, 67, 68, 69]. However, it is important to acknowledge challenges associated with SHAP values, as discussed in articles like Kumar et al. [70], which address potential limitations and issues in using SHAP values for feature importance measures. These additional articles further contribute to the understanding and application of Shapley values in different domains, including cooperative game theory, machine learning, and model explanations.

Furthermore, research in cooperative game theory has highlighted the significance of Shapley values. Studies by Littlechild [71], Kalai and Samet [72], Hart and Mas-Colell [73], Rozemberczki et al. [74], and Merrick et al. [75] contribute to the understanding and application of Shapley values in interpreting model predictions, feature importance, and cooperative game scenarios.

Explainable Artificial Intelligence (XAI) has emerged as a crucial area of research, aiming to provide transparency and interpretability to complex machine learning models. Researchers have explored various concepts, methodologies, and challenges in the pursuit of responsible AI[76, 77, 78, 79]. The DARPA XAI program [80] and articles like Arrieta et al. [81], Das et al. [82], and Van et al. [83] have contributed significantly to the understanding and advancement of XAI. They discuss taxonomies, opportunities, challenges, and approaches for achieving explainability in different domains, such as medical image analysis, clinical decision support systems, and user experiences.

Furthermore, studies by Adadi and Berrada [84], Tjoa et al. [85], and Langer et al. [86] emphasize stakeholder perspectives, interdisciplinary XAI research, and the importance of user-centered explanations in XAI. Antoniadi et al. [87] and Liao et al. [88] delve into challenges and opportunities in applying XAI to clinical decision support systems and human-centered design principles, respectively.

The survey by Vcyras and Geist [89] investigates argumentative XAI, while Saeed et al. [90] and Van et al. [91] provide systematic evaluations and comparisons of XAI methods. Wolf et al. [92] focus on scenario-based XAI design, and Paez et al. [93] introduce the pragmatic turn in XAI research.

Moreover, Minh et al. [94] present a comprehensive review of XAI, covering various techniques and applications. Schlegel et al. [77] and Rojat et al. [95] discuss XAI methods for time series analysis, while Machlev et al. [96] explore XAI techniques in energy and power systems. Additionally, the article by Kenny et al. [97] investigates post-hoc explanations-by-example and their impact on user studies.

These articles collectively contribute to the understanding and advancement of XAI, offering insights into its methodologies, challenges, and opportunities across various domains.

In conclusion, the achievements in the fields of investment analysis and environmental science have been significant and multifaceted. Researchers have made notable contributions by investigating key factors influencing investment attractiveness, analyzing air quality index, understanding regional disparities, advancing time series prediction techniques, and enhancing model interpretability. The current thesis builds upon this foundation by employing various statistical methods, machine learning models, and explainable AI techniques, furthering our understanding of these fields and providing practical applications for investors, policymakers, and researchers.

Goals and tasks of the thesis

The primary goal of this thesis is to contribute to knowledge advancement in the fields of economics and ecology, with a focus on China and ASEAN5 countries. The aim is to provide practical insights that can inform decision-making processes, facilitate economic development, promote environmental sustainability, and aid in policy formulation. To achieve this overarching goal, the following goals and corresponding tasks have been identified:

1. Assessing Investment Attractiveness: Utilize multiple linear regression to assess the investment attractiveness of China and ASEAN5 countries. Employ stepwise regression to determine the most significant factors influencing investment attractiveness. Provide insights for policymakers, businesses, and investors to make informed decisions regarding resource allocation and investment strategies.

2. Cluster Regression and Regional Analysis: Employ cluster regression techniques to divide China into distinct regional groups based on characteristics. Conduct separate regression analyses for each group to identify region-specific factors affecting economic development and environmental conditions. Facilitate tailored policymaking and resource allocation strategies to address unique challenges and opportunities within different regions.

3. Analyzing Air Quality Index: Apply stepwise regression to analyze the air quality index. Identify the key determinants and factors influencing air pollution levels. Assist policymakers and environmental agencies in formulating targeted interventions and policies to improve air quality.

4. Time Series Prediction of Air Quality: Utilize advanced neural network models such as ANN, RNN, LSTM, GRU, BiRNN, BiLSTM, and BiGRU to simulate and predict air quality in a time series context. Improve forecasting accuracy by capturing complex temporal patterns effectively. Enable proactive decision-making and targeted interventions for air pollution control and management.

5. Ensemble Modeling for Air Quality Prediction: Utilize ensemble models, including XGBoost, LightGBM, and CatBoost, to simulate and predict air quality over time. Compare the performance of these models to identify the most effective approach for air quality forecasting. Assist stakeholders in selecting appropriate techniques for accurate and reliable predictions.

6. Black Box Model Analysis using SHAP Values: Analyze the black box model using SHAP values in Explainable AI. Examine the factors contributing to model predictions and assess their impact. Enhance transparency and interpretability of the model, aiding stakeholders in responsible decision-making and understanding the underlying mechanisms.

By undertaking these main tasks, the thesis aims to achieve its overarching goal of advancing knowledge, providing practical insights, and supporting evidence-based decision-making in economics and ecology, particularly in relation to investment attractiveness and air quality in China and ASEAN countries.

Scientific novelty

This thesis presents a novel and comprehensive approach to analyzing investment attractiveness and air quality, contributing to the fields of investment analysis and environmental science. It offers scientific novelty in several aspects by integrating multiple analytical techniques and applying them innovatively. By combining multiple linear regression, cluster regression, and various machine learning models, this study provides a holistic understanding of investment attractiveness and air quality index, enhancing our understanding of these complex phenomena. The application of stepwise regression adds a novel dimension to the field of investment attractiveness research, identifying the most significant factors influencing investment attractiveness in China and ASEAN5 countries. Moreover, the utilization of cluster regression divides China into distinct groups for regional analysis, providing valuable insights into the diverse factors impacting investment attractiveness within different areas. In terms of air quality prediction, this thesis explores time series forecasting using a range of neural network models such as ANN, RNN, LSTM, GRU, BiRNN, BiLSTM, and BiGRU, contributing new insights to the existing literature. Additionally, the comparative analysis of XGBoost, LightGBM, and CatBoost models offers valuable insights into their suitability for air quality prediction tasks. Finally, this research employs explainable AI techniques, specifically SHAP value analysis, to interpret the black box models used in the research, enhancing transparency and understanding. Overall, this thesis contributes scientific novelty through its integration of multiple analytical techniques, innovative applications of regression models, exploration of neural network models, comparative analysis of boosting models, and employment of explainable AI techniques, advancing our understanding of investment attractiveness and air quality and providing significant implications for decision-making in relevant fields.

Theoretical and practical significance

Theoretical Significance:

From a theoretical perspective, this research significantly contributes to the existing knowledge in several areas. Firstly, in the analysis of investment attractiveness, the study expands our understanding of the factors driving investments in China and ASEAN5 countries. By examining variables such as per capita income, fixed assets, construction activities, and GDP per capita, this research provides insights into the complex dynamics that influence investment decisions. These findings contribute to the theoretical understanding of regional development and help establish a foundation for future research in investment attraction strategies.

Secondly, the research on air quality variations offers valuable insights into the factors influencing air pollution levels. By identifying significant variables like SO2 and NO2 through regression analysis, the study advances our understanding of the environmental determinants impacting air quality. The findings provide theoretical insights into the dynamics of air pollution, contributing to the body of knowledge on environmental management and public health.

Furthermore, the evaluation of different predictive modeling techniques for air quality prediction enhances our theoretical understanding of their effectiveness and efficiency. By comparing models such as BiRNN and LightGBM, this research contributes to the field of predictive modeling by providing empirical evidence on the performance of these models in capturing the complexity of air quality data. These findings deepen our understanding of the capabilities and limitations of various machine learning algorithms, further enriching the theoretical foundation of air quality prediction research.

Practical Significance:

The practical significance of this research lies in its implications for policymakers, businesses, and investors. The findings offer valuable insights that can inform decision-making processes and guide actions in real-world scenarios.

Firstly, the analysis of investment attractiveness provides practical guidance for policymakers seeking to attract investments and foster economic growth. By identifying key factors such as per capita income, fixed assets, and construction activities, policymakers can tailor their strategies to create an attractive investment environment. This knowledge can help direct resources effectively and implement targeted policies to encourage business development and improve regional economies.

Secondly, the research on air quality variations holds practical importance for environmental management and policymaking. Understanding the factors influencing air pollution levels, such as SO2 and NO2, allows policymakers to develop evidence-based interventions and regulations. By targeting these specific pollutants, they can implement more effective pollution control measures, enhance air quality, and safeguard public health.

Furthermore, the evaluation of predictive modeling techniques for air quality prediction has practical implications for stakeholders involved in air pollution monitoring and management. The identification of models with higher accuracy and efficiency, such as BiRNN and LightGBM, provides practical guidance for selecting the most suitable approach for air quality forecasting. This empowers stakeholders to make timely decisions, take proactive measures to mitigate pollution, and allocate resources efficiently.

Overall, the practical significance of this research lies in its ability to inform policymakers, businesses, and investors about effective strategies and actions to enhance investment attractiveness, improve air quality, and drive sustainable development. The insights gained from this study can guide decision-making processes, optimize resource allocation, and contribute to positive environmental and economic outcomes in real-world contexts.

Paper Structure and Chapter Arrangement

This paper is organized into five chapters, each exploring different aspects of investment attractiveness and air quality evaluation using various statistical methods.

Chapter 1: Analyzing Investment Attractiveness using Multiple Linear Regression: A Comprehensive Study on Identifying Key Factors for Sustainable Economic Development. In this chapter, the multiple linear regression method is employed to simulate investment attractiveness. Additionally, the stepwise regression technique is utilized to identify the key factors that impact investment attractiveness. The chapter provides a comprehensive analysis of these factors.

Chapter 2: Exploring Investment Attractiveness: An In-depth Analysis Using Cluster Analysis Method. This chapter delves into exploring investment attractiveness in greater detail by utilizing the Cluster Analysis Method. Based on this method, the entire Chinese region is segmented into four distinct groups, allowing for modeling and data analysis within each group. The chapter presents an in-depth examination of investment attractiveness within these clusters.

Chapter 3: Analyzing and Simulating Air Quality Index using Stepwise Regression: Exploring Trends and Evaluating Fit. This chapter focuses on the use of stepwise regression to analyze and simulate the Air Quality Index (AQI). It provides an overview of the significance of studying air quality, explains the stepwise regression methodology, and describes the process of selecting influential variables. The chapter also discusses the exploration of trends and evaluates the goodness of fit of the model.

Chapter 4: Deep Learning Methods for Air Quality Evaluation System. In this chapter, deep learning methods are introduced, including models such as ANN, RNN, LSTM, GRU, BiRNN, BiLSTM, and BiGRU. Experimental simulations of air quality, specifically PM2.5, are conducted using these models. The performance of each model is evaluated from various perspectives, providing insights into the quality and effectiveness of these seven models.

Chapter 5: Ensemble Learning Methods for Air Quality Evaluation System. This chapter focuses on ensemble learning methods, namely XGBoost, LightGBM, and CatBoost. Additionally, it introduces interpreting outcomes through SHAP-based explanations. Simulations and predictions of air quality (PM2.5) data are conducted using the three models mentioned. Furthermore, the chapter analyzes the influencing factors based on the SHAP approach, which enables interpretation of the originally black-box model.

By structuring the paper into these five chapters, the research covers a comprehensive investigation into investment attractiveness and air quality evaluation. The utilization of multiple statistical approaches allows for a thorough exploration of various factors and techniques, contributing to a well-rounded analysis of these important domains.

Results submitted for defense

1. Statistical modeling conducted to assess the investment attractiveness of China regions and ASEAN countries. Factors influencing investment attractiveness in China regions and ASEAN countries analyzed using stepwise regression. Cluster analysis performed to explore investment attractiveness variations within different regions of China.

2. Statistical modeling employed to examine the dynamics of air quality in China regions. Factors affecting air quality identified through stepwise regression analysis.

3. Comparative evaluation of deep learning algorithms (ANN, RNN, BiRNN, LSTM, BiLSTM, GRU, BiGRU) for time series prediction of air quality.

4. Comparative assessment of ensemble learning algorithms (LightGBM, CatBoost, XGBoost) for time series forecasting of air quality. Factor analysis of black box models using SHAP values in Explainable AI (XAI).

Main scientific achievements

The **main scientific achievements** obtained during the dissertation research include the following:

1. The study focused on investigating investment attractiveness. Initially, multiple regression analysis was employed to construct models for assessing investment

attractiveness ([98], p. 3). Subsequently, a stepwise regression method was utilized to identify and incorporate the most influential factors into each model ([98], p. 4). To summarize the outcomes of each model comprehensively, a final analysis was conducted to evaluate the impact of these identified factors on investment attractiveness ([98], p. 5). Finally, based on the findings, the authors proposed a list of the most crucial determinants influencing investment attractiveness ([98], p. 6). The detailed results of this research are documented in the publication by the authors [98]. The researchers were actively engaged in various stages of the study, encompassing data collection, model development, result analysis, literature review, result interpretation, and manuscript composition.

2. Building upon the foundational research cited in [99], an expanded dataset was utilized for a thorough classification analysis of investment attractiveness. Initially, the regions under study were categorized into four distinct groups based on their levels of investment appeal ([99], p.4). Subsequently, comparisons of the investment attractiveness within each group were conducted to assess variations and nuances ([99], p.5). Individual models were then tailored for each group, followed by detailed analyses to elucidate specific trends and insights ([99], pp.8–11). The study's outcomes revealed that the strategies employed to attract investments vary significantly across different regions and groups ([99], p.12). The comprehensive findings stemming from this investigation are documented in the publication by the authors [99]. The researchers actively participated in all aspects of the research process, encompassing data collection, review of relevant literature, result interpretation, and manuscript composition.

3. The primary objective of this study was to identify and evaluate the key determinants influencing the dependent variable. The research findings are exemplified through a case study focusing on air quality conditions in China ([100], p.2). Specifically, the analysis delves into aspects such as model characteristics, quality assessment, parameter validation, and residual diagnostics ([100], pp.3–4). Subsequently, in pursuit of refining the predictive accuracy of the model, the backward elimination stepwise regression technique was employed to derive the final refined model, demonstrated within the context of China's air quality

scenario ([100], p.6). Concurrently, a detailed examination of each analytical step was conducted to provide insights into the outcomes regarding air quality dynamics in China ([100], p.7).The established model was then utilized to forecast the annual Air Quality Index (AQI) for 31 provincial capital cities in China spanning the years 2013 to 2019. The resulting forecasts were substantiated by actual air quality data from China ([100], p.8). These comprehensive findings were disseminated in the publication by the author [100]. The researcher played an active role in the acquisition of these results, encompassing material collection, literature data analysis, and result interpretation.

4. The primary objective of this research is to assess the long-term air quality trends in China by employing multinomial logistic regression methods based on the Air Quality Index (AQI) and the Air Quality Comprehensive Index (AQCI). Two distinct models were developed, each utilizing different dependent variables—AQI and AQCI—while maintaining consistent control variables such as gross domestic product (GDP) and major pollutants ([101], p.4). Specifically, the principal pollutants considered in the analysis are linked to one or more of six pollutant factors: O3, PM2.5, PM10, NO2, SO2, and CO ([101], p.6). Ensuring the quality and validity of the models is paramount and forms an essential component of the analytical process ([101], p.7). The outcomes, published in the referenced study [101], are elucidated using authentic air quality data sourced from China. The author was actively engaged in the acquisition of these findings, encompassing material gathering, literature review, and result interpretation.

5. This paper conducts an in-depth analysis of time series prediction models utilized for forecasting air quality. The research focuses on the identification and evaluation of predictive models suitable for environmental analysis, encompassing prominent algorithmic frameworks such as neural networks and ensemble models ([102], pp.4–6). The effectiveness and performance of these models are assessed utilizing key metrics including mean absolute error (MAE), root mean square error (RMSE), and R-squared values ([102], pp.7–9). The findings indicate that neural networks and ensemble models exhibit robust capabilities in predicting air quality time series data reliably ([102], p.12). These results

have been formally disseminated in the publication by the author [102]. The researcher played an active role in various stages of the research process, involving material collection, literature data analysis, result interpretation, and manuscript composition.

6. This scholarly article presents a comprehensive examination of ShapTime, an eXplainable Artificial Intelligence (XAI) methodology hinged on Shapley values, specifically engineered to enhance the interpretability and efficacy of time series forecasting by delving into intricate temporal dynamics. Key innovations associated with ShapTime encompass its capacity to offer stable explanations, thereby reflecting the intrinsic significance of time itself and rendering it more adept for time series prediction compared to conventional XAI techniques ([103], p.8). Furthermore, the study introduces a pragmatic application framework within XAI wherein the elucidated outcomes serve as guiding principles to enhance forecasting precision, setting it apart from prior studies that solely utilized XAI results as demonstrations of novelty ([103], p.10). Notably, across five diverse real-world datasets, ShapTime showcased notable average performance enhancements in Boosting, RNN-based models, and Bi-RNN models, culminating in improvements of 18%, 20%, and 35%, respectively ([103], p.13). These research findings have been formally documented and disseminated in the referenced publication [103], with the author actively engaged in the meticulous processes of material gathering, literature analysis, and result interpretation, underscoring a hands-on approach to knowledge acquisition.

Chapter 1

Analyzing Investment Attractiveness using Multiple Linear Regression: A Comprehensive Study on Identifying Key Factors for Sustainable Economic Development

Discussion presented in the chapter is published in paper [98].

1.1 Groups of Research Objects

1. **Regions of China** (Provinces with low investment attractiveness have been excluded from consideration)

   - Beijing
   - Tianjin
   - Hebei
   - Shanxi
   - Inner Mongolia
   - Liaoning
   - Jilin
   - Heilongjiang
   - Shanghai
   - Jiangsu

   - Zhejiang
   - Anhui
   - Fujian
   - Jiangxi
   - Shandong
   - Henan
   - Hubei
   - Hunan
   - Guangdong
   - Guangxi
   - Chongqing
   - Sichuan
   - Guizhou
   - Yunnan
   - Shaanxi
   - Xinjiang

2. **ASEAN-5**

   - Indonesia
   - Malaysia
   - Singapore
   - Thailand
   - Philippines

1.2 Data and Methods

This section introduces the principles, assumptions, and application scenarios of the multiple linear regression model. It explains how to apply this method to study the investment attractiveness of China and the five ASEAN countries, providing calculation formulas and parameter estimation methods.

1.2.1 Data Source and Collection

Data on Investment Attractiveness of Chinese Regions: The data on investment attractiveness of various regions in China is collected from the China National Statistical Yearbook. This authoritative resource, issued by the National Bureau of Statistics of China, provides detailed statistical data on the economy, population, and society of all provinces, autonomous regions, and municipalities directly under the Central Government. By collecting data from 2008 to 2017, we obtain a ten-year time span for comprehensive analysis of the dynamic changes in investment attractiveness across different regions of China.

Data on Investment Attractiveness of ASEAN Countries: Data on investment attractiveness of ASEAN countries is sourced from the World Bank database. The World Bank, an international organization, offers an extensive range of economic, social, and environmental data globally. Accessing this database allows us to obtain data on the investment attractiveness of ASEAN countries from 1998 to 2014. With 16 years of data, we can analyze the investment attractiveness of ASEAN countries over the past two decades.

1.2.2 Data Cleaning and Outlier Handling

After data collection, it is necessary to perform data cleaning and outlier processing to ensure data accuracy, reliability, and remove any outliers that may distort the analysis results.

Data cleaning involves operations such as data inspection, missing value handling, and data format conversion. Thoroughly examining the collected data ensures data integrity and consistency. If any missing values are identified, appropriate measures are taken to fill or delete them. Additionally, depending on research requirements, the data may need to be converted from its original format into numerical variables suitable for multiple linear regression analysis.

Outlier processing aims to exclude extreme observations that significantly deviate from the normal range, ensuring the reliability of analysis results.

By collecting data from authoritative sources like the China National Statistical Yearbook and the World Bank database, as well as performing data cleaning and outlier processing, reliable data on the investment attractiveness of China and the five ASEAN countries can be obtained. This provides a logical and academic basis for subsequent multiple linear regression analysis.

1.2.3 The multiple linear regression method

The multiple linear regression method is an essential tool in various academic disciplines, including economics, finance, social science, and environmental science. It is based on mathematical statistics principles and is utilized to analyze the relationship between independent and dependent variables, establish a linear equation to express this relationship, and make predictions and explanations.

By assuming a linear relationship between the independent and dependent variables, the multiple linear regression model formulates a linear equation comprising regression coefficients and an intercept term. These coefficients represent the magnitude and direction of influence that independent variables have on the dependent variable, while the intercept term denotes the value of the dependent variable when all independent variables are zero. The best-fit regression coefficients are estimated by minimizing the sum of squared residuals, which measure the discrepancy between observed values and those predicted by the linear equation.

The efficacy and reliability of multiple linear regression models can be evaluated using statistical indicators, with the coefficient of determination (R?) being a commonly employed metric. R? quantifies the proportion of variance in the dependent variable that can be explained by the model. A higher R? indicates a better fit of the data. Additionally, the adjusted coefficient of determination can account for the impact of the number of independent variables on the R?, and the F statistic can assess the significance of the regression model.

Prior to conducting multiple linear regression analysis, several prerequisites and assumptions should be taken into account. Firstly, the assumption of a linear relationship expects a uniform increase or decrease between the independent and dependent variables. Secondly, the independence of independent variables is necessary, without any multicollinearity issues where one independent variable can be predicted by others. Furthermore, the error term should adhere to certain assumptions, such as independent and identically distributed errors, as well as constant variance.

To ensure that the model satisfies these prerequisites and assumptions, diagnostic tests are employed. These tests include examining residual plots, evaluating the normality and independence of residuals, among others.

In conclusion, the multiple linear regression method is a logical and academic technique used to model and analyze the relationship between independent and dependent variables. By applying this approach, researchers can gain insights into variable impact and interpretation, facilitating predictions and inferences.

1.2.4 The multiple stepwise regression method

The multiple stepwise regression method is commonly employed in research to explore and build multiple linear regression models. Unlike traditional multiple linear regression, this approach gradually eliminates independent variables to construct a more concise and explanatory model.

The primary objective of this method is to identify independent variables that significantly impact the dependent variable while excluding those with weak or insignificant explanatory power. The iterative process of removing independent variables follows pre-defined elimination criteria. Typically, these criteria are based on statistical significance levels, such as a predetermined p-value threshold for removal. Alternatively, other indicators, such as maximum adjusted coefficient of determination or non-significant F statistics, can be utilized.

Multiple stepwise regression offers several advantages. Firstly, it aids in the identification of influential variables from a pool of potential predictors, resulting in a more parsimonious model with enhanced explanatory capabilities. This assists in mitigating redundant information and improving the model's interpretability and predictive performance. Secondly, the method effectively addresses multicollinearity issues caused by high correlations among independent variables. During the stepwise deletion process, highly correlated variables are often excluded, thereby minimizing collinearity's negative impact on model outcomes. Moreover, multiple stepwise regression accommodates small sample sizes by limiting the number of independent variables, thus reducing the risk of overfitting.

Caution must be exercised when utilizing multiple stepwise regression methods. Careful consideration should be given to the selection of elimination criteria, incorporating domain knowledge and research objectives. Excessive elimination may lead to the exclusion of vital independent variables, while retaining insignificant ones may introduce noise into the model. Additionally, since multiple stepwise regression relies on a data-driven approach, the results may be subject to sample-specific influences.

In conclusion, the multiple stepwise regression method provides a means to construct compact and explanatory multiple linear regression models. By progressively removing independent variables, it facilitates the selection of a subset that displays the most relevant and significant relationships with the dependent variable, ultimately enhancing the model's explanatory power and predictive performance. However, when employing this method, careful selection of elimination criteria and thorough validation and evaluation are crucial to ensure the model's reliability and validity.

1.2.5 Constructing and analyzing models of investment attractiveness in the regions of China

Data on China's investment attractiveness shows a growing trend from 2008 to 2018 (see Figure 1).

Table 1 presents the investment volume in China's fixed assets (in billions of yuan) for the years 2008-2018.

Year	Investment Volume in Fixed Assets (100 million yuan)
2008	172828
2009	224599
2010	251684
2011	311485
2012	374695
2013	446294
2014	512021
2015	562000
2016	606466
2017	641238
2018	645675

Table 1.1: Investment Volume in China's Fixed Assets

The analysis covers the period from 2008 to 2017. For each year, a regression model is constructed to reflect the level of investment (y) based on a set of explanatory variables ($x_k, k = 1, .., 10$). The regression model is built for the entire country, considering regions as individual observations.

Factors considered:

- $\hat{y}_t$ - logarithm of the estimate of the volume of investments in the current year;

- $x_{1,t}$ - electricity consumption;
- $x_{2,t}$ - per capita income in the current year;
- $x_{3,t}$ - debt from loans provided to legal entities by credit organizations in the current year;
- $x_{4,t}$ - value of fixed assets in the current year;
- $x_{5,t}$ - expenditure on scientific research in the current year;
- $x_{6,t}$ - volume of activities in the "Construction" sector in the current year;
- $x_{7,t}$ - number of enterprises and organizations in the current year;
- $x_{8,t}$ - turnover of retail trade in the current year;
- $x_{9,t}$ - GDP per capita in the current year;
- $x_{10,t}$ - unemployment rate in the current year (in percentage);

The equation represents the regression model used to logarithm of the estimate of the volume of investments in the current year $\hat{y}_t$ in the current year based on the explanatory variables $(x_{1,t}, x_{2,t}, \cdots, x_{10,t})$.

$$\begin{aligned} \hat{y}_t &= b_0 + b_1 x_{1,t} + b_2 x_{2,t} + b_3 x_{3,t} + b_4 x_{4,t} + b_5 x_{5,t} + b_6 x_{6,t} + b_7 x_{7,t} \\ &\quad + b_8 x_{8,t} + b_9 x_{9,t} + b_{10} x_{10,t} \end{aligned} \tag{1.1}$$

where t represents the year number, with $t = 1, ..., 10$.

For each year with the number $t(t = 1, ..., 10)$, there is information available for individual regions. There are a total of 26 regions with a high level of investment.

$y_{1,t}, \ldots, y_{26,t}$ represent the investment levels in the regions.

$x_{i,1,t}, \ldots, x_{i,10,t}$; where $i = 1, \ldots, 26$, and i denotes the region number.

Based on the information for year t presented on the previous slide, a regression model (Equation 1) is constructed using the method of least squares:

$$\sum_{i=1}^{26} (y_{i,t} - b_0 - b_1 x_{i,1,t} - b_2 x_{i,2,t} - .. - b_{10} x_{i,10,t})^2 \to \min_{b_0,..,b_{10}} \tag{1.2}$$

To perform the calculations, multiple linear regression analysis was applied. Using the "Regression" tool in the "Data Analysis" add-in in MS Excel, we will conduct a regression analysis on the available values of the column vector Y and X.

Multiple regression model for the year 2017

The coefficient estimates, calculated using the regression tool in the data analysis add-in of MS Excel, are provided for the observed values in Table 1.2 (standard errors of the coefficients are indicated in parentheses). The parameter estimates for the regression equation in the year 2017 are as follows:

b10(c.o.)	b9(c.o.)	b8(c.o.)	b7(c.o.)	b6(c.o.)	b5(c.o.)
90.12045361	0.757670787	-0.074390939	0.002579218	0.678321663	-0.001846309
(1645.726812)	(0.426035564)	(8.61877992)	(0.008987614)	(0.302921691)	(0.001176357)

Table 1.2: The parameter estimates for the regression equation in the year 2017

b4(c.o.)	b3(c.o.)	b2(c.o)	b1(c.o.)	b0(c.o.)
0.760781591	-5.183897966	-0.399105119	-1.873479417	9312.257203
(0.457066773)	(3.137349872)	(0.15744744)	(1.834540235)	(8232.963495)

The model is expressed as follows:

$$\begin{aligned}\hat{y}_t = {} & 9312.2572 - 1.8734 \cdot x_{1,t} - 0.3991 \cdot x_{2,t} - 5.1838 \cdot x_{3,t} + 0.7607 \cdot x_{4,t} \\ & - 0.00184 \cdot x_{5,t} + 0.6783 \cdot x_{6,t} + 0.00257 \cdot x_{7,t} - 0.07439 \cdot x_{8,t} \\ & + 0.7576 \cdot x_{9,t} + 90.1204 \cdot x_{10,t}\end{aligned} \tag{1.3}$$

Next, we will verify that the overall form of the studied functional relationships is successfully determined. This involves assessing the quality and adequacy of the constructed model based on empirical data.

The main criteria for evaluating the quality of multiple regression are the coefficient of determination R^2 , the multiple correlation coefficient R, and the adjusted coefficient of determination R^2_{adj}.

Subsequently, models are constructed to determine the values of R^2, R and R^2_{adj} as shown in Table 1.3.

$R^2 = 0,935$, indicating that approximately 94% of the variation in the studied variable y can be "explained" by the combination of factors included in the model.

R	0.967293678
R^2	0.93565706
R^2_{adj}	0.892761767

Table 1.3

Taking into account the penalty imposed for the large number of explanatory variables, $R^2_{\text{adj}} = 0,892$, meaning that the regression equation accounts for 89% of the variance in the dependent variable (within the observed values of y). The multiple correlation coefficient $R = 0,976$, which is very close to 1, indicates a strong relationship between y and the set of predictor variables $x_1, \cdots, x_{10}$.

The classical approach to testing the adequacy of econometric models is the F-test, which involves assessing the significance of the regression equation based on the Fisher criterion and comparing this assessment with the critical value of the test. The critical value depends on the confidence interval and the degrees of freedom k; $n-k$. If the computed value exceeds the critical value, then the model is considered suitable for analysis at the chosen level of significance.

For our model, the value of $F = 21,8125$; $n = 26$; $k = 10$. Let us set the significance level α, $\alpha = 0,05$ as is traditionally done in statistical research.We calculate the critical value F_{crit} as $F_{(0,05;10;26-10-1)} = 2,543$. Since $F > F_{\text{crit}}$, the model is statistically significant at the $\alpha = 0,05$. Additionally, using the Excel report, we can evaluate the significance of the equation by comparing the p-value(F) with α. If $p - \text{value}(F) < \alpha$, then the regression model is considered statistically significant at the specified significance level. For the analyzed model, $p-\text{value}(F) = 3,72 \cdot 10^{-07}(< 0,05))$.

Therefore, the regression equation is statistically significant at $\alpha = 0,05$, which confirms the results of the F-test.

Evaluation of the quality of regression equation parameters. We compare the absolute values of the computed t-statistics of the obtained regression model coefficients with the tabulated values. A parameter estimate is considered significant in all cases when the absolute computed (observed) value exceeds the tabulated value.

For a given significance level $\alpha = 0,05$ and degrees of freedom $df = 26-10-1 = 15$ the theoretical (tabulated) value of the t-statistic is $t_{\text{table}} = 2,131$.

Table 1.5 presents the computed values of the t-statistics for the respective coefficients of the regression model. It is evident that only the estimates for α_2 and α_6

are statistically significant at the 5% significance level. This table also includes the p-values for the observed t-statistics. The p-value (t) represents the critical value of the significance level α for the current t-statistic magnitude. Thus, if the p-value is smaller than the specified α, the coefficient is considered statistically significant. In terms of the p-value, the evaluation of the significance of the parameters in the constructed regression equation leads to the same results as the t-test.

	t Stat	P-value
b0	1.131094193	0.275776099
b1	-1.021225581	0.323336646
b2	-2.534846664	0.0228745
b3	-1.652317458	0.119242738
b4	1.664486759	0.11675925

Table 1.4: t-statistics of the parameters of the regression equation for 2017.

	t Stat	P-value
b5	-1.569513503	0.13737914
b6	2.239264087	0.040718589
b7	0.286974677	0.778057886
b8	-0.008631261	0.993227085
b9	1.778421455	0.095605091
b10	0.054760275	0.957052161

Table 1.5: t-statistics of the parameters of the regression equation for 2017.

Thus, the constructed model demonstrates high approximation quality. However, out of the 11 parameter estimates, only 2 are found to be significant. This outcome is likely due to the presence of multiple linear factors in the model. Considering the relatively large number of explanatory variables and the relatively small sample size, it is reasonable to assume that some indicators may duplicate each other. Thus, it is necessary to adjust the set of factors by using a stepwise regression method. We will sequentially exclude variables corresponding to statistically insignificant coefficient estimates until all of them become significant at the given α level.

For the "screening" indicator, we will consider the p-value. The variable associated with the coefficient having the highest p-value will be removed from the model. Afterward, using the transformed set of explanatory features, we will rebuild the regression equation for the existing y.

Referring to Table 1.5, let us identify the largest p-value. The coefficient b_8 for the variable x_8 has the smallest absolute t-statistic value and, consequently, the largest p-value. Therefore, we will exclude x_8 the model. As a result, the matrix of original data will take the form:

$X = (x_{1i}, x_{2i}, \cdots, x_{7i}, x_{9i}, x_{10i})^\top$,$i = 1, \cdots, 26$.

For $y = (y_1, \cdots, y_{26})^\top$ and the new X, we use MS Excel to recalculate the parameter estimates of the regression equation.

The coefficient of determination and the coefficient of multiple correlation remain unchanged, while R^2_{adj} increases, indicating that the remote explanatory variable is statistically insignificant. The updated model now has an adjusted coefficient of determination $R^2_{adj} = 0.899$, indicating a reasonably high level of quality. The significance level of the computed Fisher criterion value is lower compared to the previous equation. With a $P-$value$(F) = 6.97669 \cdot 10^{-8}$, the transformed regression model is considered statistically significant at $\alpha = 0,05$.

The coefficient estimates:

b10(c.o.)	b9(c.eo.)	b7(c.o.)	b6(c.o.)	b5(c.o.)
91.62080512	0.75698959	0.002606575	0.678749549	-0.001850222
(1584.558496)	(0.405368437)	(0.008143127)	(0.289349084)	(0.001051)

Table 1.6: The parameter estimates of the regression equation for the year 2017 without x_8

b4(c.o.)	b3(c.o.)	b2(c.o.)	b1(c.o.)	b0(c.o.)
0.760061396	-5.18729215	-0.400113466	-1.869145659	9322.593323
(0.435117166)	(3.013776515)	(0.102202319)	(1.708466631)	(7886.774517)

Evaluation of the quality of estimates based on the coefficients of the new equation. The observed t-statistic values and their corresponding p-values are presented in Table 1.8.

	t Stat	P-value
b0	1.182053995	0.254454547
b1	-1.094048678	0.29013139
b2	-3.914915737	0.001234375
b3	-1.721193368	0.104490736
b4	1.746797082	0.099837148

Table 1.7: The t-statistics after removing x_8 for the year 2017.

	t Stat	P-value
b5	-1.76043987	0.097432206
b6	2.345780882	0.032203352
b7	0.320095135	0.753038323
b9	1.867411275	0.080274135
b10	0.05782103	0.954607156

Table 1.8: The t-statistics after removing x_8 for the year 2017.

Considering the constant term, there are 10 coefficients. Based on the estimated values and in accordance with the results of the t-test, only 2 of them are significant: b_2 and b_6. The p-value reaches its maximum value at $t = 0,0578$, calculated for the estimate b_{10}. Here, b_{10} represents the coefficient for the explanatory variable x_{10}. As a result, x_{10} is removed from the set of factors. Thus, $X = (x_{1i}, x_{2i}, \cdots, x_{7i}, x_{9i})^\top$, where $i = 1, \cdots, 26$.

The regression statistics of the model, obtained after removing the factor x_{10} , exhibit only minor differences compared to the statistics of the previous model (see Table 1.9):

R	0.9672865
R^2	0.935643
R^2_{adj}	0.905357

Table 1.9

The observed F-statistic (F = 37.1309) exceeds the values corresponding to previous equations, with a , $P-\text{value}(F) = 1.95853 \cdot 10^{-9}$, indicating a high level of approximation quality.

The parameter estimates of the model (see Table 1.10) are predominantly statistically insignificant. The estimate of the coefficient b_7 for the explanatory variable x_7 has the highest $P-\text{value}(F) = 0,7458$. Thus, we remove the factor x_7 as it is not statistically significant in the analyzed model. Its lack of statistical significance is further confirmed by qualitative assessments of the equations' R^2 and R values obtained after exclusion, which remain almost at the same level. The adjusted R^2 increases due to the reduction in the number of explanatory variables ($R^2_{\text{adj}} = 0,910$). The p-value (F) decreases, and the confidence interval at which the model is statistically significant at $\alpha = 0,05$ widens.

By excluding the least significant explanatory variables x_1 and x_4 from the models built in the following two steps, we obtain a model where all parameter estimates are statistically significant at $\alpha = 0,05$. The $P-\text{value(F)} = 2.8909 \cdot 10^{-10}$ indicates that the model is statistically significant.

The model takes the following form:

		c.o.	t Stat	P-value
b0	10034.49788	3991.129234	2.51420019	0.021658637
b1	-1.610506826	1.429250709	-1.12681898	0.274617446
b2	-0.402520397	0.094263852	-4.270145875	0.000460606
b3	-5.433737956	2.681372429	-2.026476403	0.057795086
b4	0.734552101	0.397333472	1.848704307	0.0809958
b5	-0.001708463	0.000903599	-1.8907318	0.074869
b6	0.743294813	0.198865735	3.737671606	0.001506765
b9	0.769269896	0.376296307	2.044319548	0.055832772

Table 1.10: The parameter estimates of the regression equation for the year 2017 without x_8,x_{10} ,x_7

$$\begin{aligned}\hat{y}_t \underset{(t-Stat)}{=} & \underset{(2.8720)}{10712.1937} - \underset{(-5.2899)}{0.4501 \cdot x_{2,t}} - \underset{(-2.4570)}{6.7041 \cdot x_{3,t}} - \underset{(-3.1611)}{0.002625 \cdot x_{5,t}} \\ & + \underset{(3.6163)}{0.7599 \cdot x_{6,t}} + \underset{(4.9829)}{1.2368 \cdot x_{9,t}}\end{aligned} \tag{1.4}$$

The overall quality of the adopted regression equation is quite high, as confirmed by the values of the indicators R^2, R^2_{adj}, and R shown in Table 1.11.

Table 1.11: Dynamic Quality Metrics during Variable Selection in 2017

Model	**Multiple R**	**R-Squared**	**Norm. R-Squared**	**F-Value**	**p-value(F)**	AIC	p(RNT)	p(HT)
1,2,3,4,5,6,7,**8**,9,10	0.967	0.935	0.892	21.81	3.7254E-07	519.3	0.8756	0.3478
1,2,3,4,5,6,7,9,**10**	0.967	0.935	0.899	25.85	6.9766E-08	517.3	0.8771	0.3478
1,2,3,4,5,6,**7**,9	0.967	0.935	0.905	30.89	1.1947E-08	515.3	0.8831	0.3479
1,2,3,4,5,6,9	0.967	0.935	0.910	37.13	1.9585E-09	513.5	0.8059	0.3757
2,3,**4**,5,6,9	0.964	0.930	0.908	42.50	5.1728E-10	**513.3**	0.6676	0.8337
2,3,5,6,9	0.958	0.919	0.899	45.69	2.8909E-10	515.1	0.46	0.6470

The model with the lowest AIC:

$$\begin{aligned}\hat{y}_t =& 8484.5941 - 0.375 \cdot x_{2,t} - 5.4173 \cdot x_{3,t} + 0.6973 \cdot x_{4,t} \\ & - 0.0019 \cdot x_{5,t} + 0.7458 \cdot x_{6,t} + 0.7238 \cdot x_{9,t}\end{aligned} \tag{1.5}$$

The presented model (4) characterizes 89.9% of the variance in the initial values of y_i , where $i = 1, \cdots, 26$ represents the observation number (region), taking into account penalties for each new explanatory variable. The coefficient of multiple correlation is $R = 0,958$, indicating a strong linear relationship between the outcome variable and the set of factors characterizing it. The model has been found statistically significant based on the Fisher criterion at a significance level of $\alpha = 0,05$ with $F = 45.6923$. For the computed F-statistic value, the $P - \text{value}(F) = 2.89 \cdot 10^{-10}$.

The diagrams of the theoretical values y_i and the actual values y for the observation numbers $i = 1, \cdots, 26$ (see Fig. 1.1) provide compelling evidence of the good quality of approximation between the original values of y and the studied model.

By using the Durbin-Watson criterion, it is possible to determine whether these residuals exhibit autocorrelation or not. The estimated d-statistic (DW) takes the following form:

$$DW = \frac{\sum_{i=2}^{n} \left(\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1}\right)^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}.$$

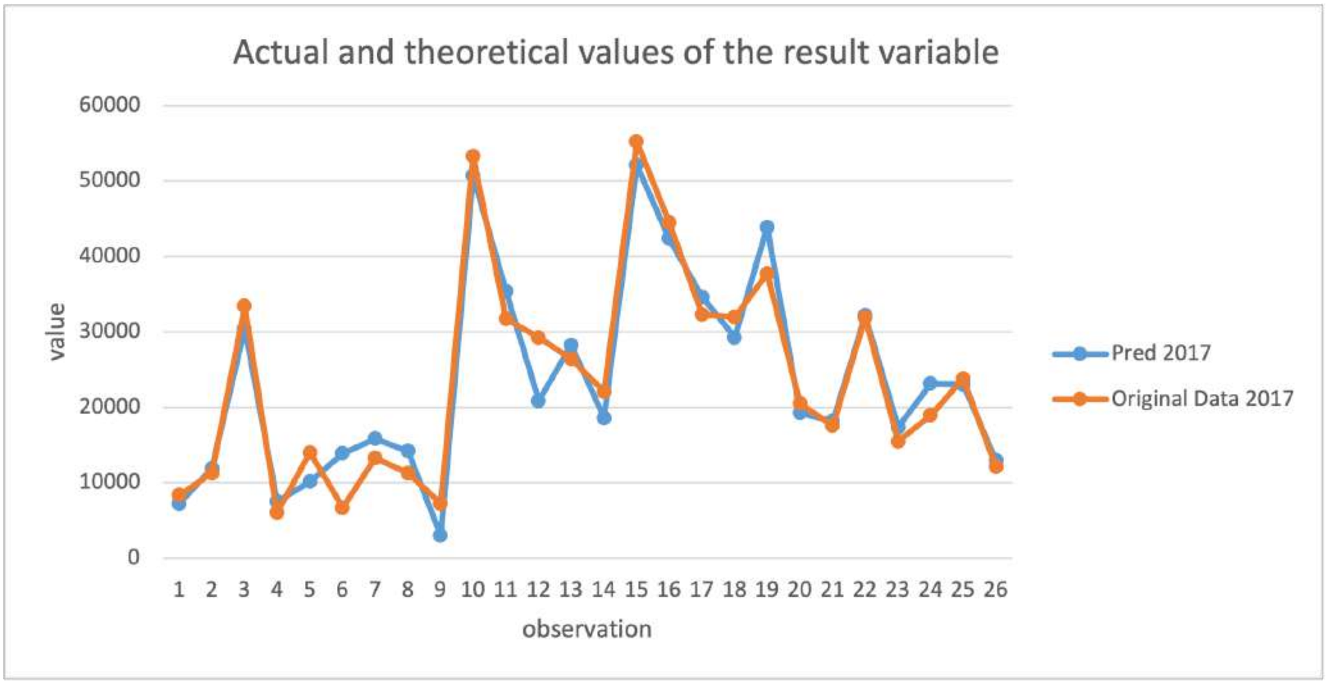


Figure 1.1

The essence of the criterion lies in comparing the empirical value of DW with the tabulated lower (d_L) and upper (d_U) statistics, which depend on the sample size, significance level, and number of explanatory variables in the model. If $DW > 2$, we consider the adjusted value $DW' = 4 - DW$.

According to the conditions of regression analysis ($\alpha = 0,05, n = 26, k = 15$), the tabulated values are $d_l = 0,256$ and $d_U = 3,179$. There is no autocorrelation if both conditions hold: $DW > d_U$ and $4 - d_U > DW$. In applied statistical data analysis, it is generally recognized that when the statistical indicators of the Durbin-Watson criterion range from 1.5 to 2.5, the regression model can be considered adequate. In our equation, $DW' = 2,596$, and the observed value falls into the "zone of uncertainty" where $d_L < DW' < d_U$.

Let's calculate the average relative approximation error $E_{\text{rel.}}$in order to assess the accuracy of the regression model, using the following formula:

$$E_{\text{rel.}} = \frac{1}{n} \sum_{i=1}^{n} \frac{|\hat{\varepsilon}_i|}{\ln y_i} \cdot 100\%.$$

For the constructed regression equation, this value is $E_{\text{rel.}} = 1,75\%$,, indicating an extremely small model error (around 2%).

Thus, the developed regression model in the form of equation (1.4), considering its high-quality indicators and minimal approximation error, demonstrates statistically significant influence of the set of explanatory factors on the value of the dependent variable. It is also suitable for further analysis of probabilistic dependencies between lny and the variables x_2, x_3, x_5, x_6, x_9 within the scope of the year 2017.

Multiple regression model for 2016

Based on the data for the year 2016 provided by the National Bureau of Statistics (China), a computational table has been constructed to determine the parameters of the regression equation, which takes the following form:

$$\begin{aligned} y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}, x_{10i}), \\ i = 1..26. \end{aligned} \tag{1.6}$$

where: i represents the number of the region under investigation;
y_i - logarithm of the estimate of the volume of investments in the current year in the i-th region;
x_{1i} - electricity consumption in the i-th region;
x_{2i} - per capita income in the current year in the i-th region;
x_{3i} - debt from loans provided to legal entities by credit organizations in the current year in the i-th region;
x_{4i} - value of fixed assets in the current year in the i-th region;
x_{5i} - expenditure on scientific research in the current year in the i-th region;
x_{6i} - volume of activities in the "Construction" sector in the current year in the i-th region;
x_{7i} - number of enterprises and organizations in the current year in the i-th region;
x_{8i} - turnover of retail trade in the current year in the i-th region;
x_{9i} - GDP per capita in the current year in the i-th region;
x_{10i} - unemployment rate in the current year (in percentage) in the i-th region;

The coefficient estimates are calculated using the regression tool in MS Excel's data analysis add-in for the observed values $y = (y_1, ..., y_26)^\top$ and $X = (x_{1i}, x_{2i}, \cdots, x_{10i})^\top$, where $i = 1, \cdots, 26$. These estimates are presented in the form shown in Table 1.12:

The overall model quality criteria indicate a high level of its approximation ability.

	Coefficients	c.o.
b0	3262.169924	4832.230643
b1	-0.492576567	1.746298859
b2	-0.284674564	0.137878968
b3	-2.647672589	2.925816588
b4	1.36612506	0.335356386
b5	-0.000111808	7.46915E-05
b6	8.62503E-05	3.41244E-05
b7	0.002335237	0.008644545
b8	-0.475518489	7.629315583
b9	-0.040903966	0.284922846
b10	915.815172	1211.023343

Table 1.12: Estimates of the parameters of the regression equation for the year 2016.

The coefficient of multiple determination is $R^2 = 0,937$, and after adjusting for the penalty imposed by the number of explanatory variables, the adjusted coefficient of determination becomes $R^2_{adj} = 0,906$. Additionally, the multiple correlation coefficient approaches unity with $R = 0,968$. The observed value of the F-statistic is $F = 30.066$. At a given significance level of $\alpha = 0,05$, the critical value is $F_{0,05;10;15} = 2,544$. Thus, the model is statistically significant and explains almost the entire variance of the outcome variable. However, despite the overall good quality of the regression model, the parameter estimates for the explanatory variables are not considered statistically significant for the chosen level of α. (See Table 1.13)

	t Stat	P-value
b0	0.675085724	0.507351564
b1	-0.282068882	0.780788875
b2	-2.064669959	0.052164857
b3	-0.904934574	0.376272554
b4	4.073651542	0.000592215
b5	-1.496936962	0.150025462
b6	2.527527177	0.02001929
b7	0.270139899	0.789820503
b8	-0.062327804	0.950920479
b9	-0.143561551	0.887283334
b10	0.756232469	0.458325579

Table 1.13: t-statistics for the year 2016

Let's proceed with the stepwise exclusion methodology for removing statistically insignificant factors from the regression equation. Based on the results obtained in the second stage, a model is constructed without x_8 ($b_8 = 0$). The new model has almost the same quality as the previous one, but it is statistically more significant according to the Fisher's F-test.

The dynamics of the changes in the indicators R^2,R, R^2_{adj}, and F during the sequential removal of insignificant factors from the multiple regression model are presented in Table 1.14.

Explanatory variables	R	R^2	$R_{\text{adj}}{}^2$	$\text{F}_{\text{obs}}.$	p-value(F)	AIC	p(RNT)	p(HT)
x1,x2,x3,x4,x5,x6,x7,***x*9**,x10	0.9592	0.9201	0.8751	20.4764	3.74E-07	519.6	0.6831	0.4306
***x*1**,x2,x3,x4,x5,x6,x7,x10	0.9591	0.9199	0.8823	24.4319	7.27E-08	517.6	0.6685	0.4313
x2,x3,x4,x5,x6,***x*7**,x10	0.9589	0.9196	0.8883	29.4284	1.31E-08	515.7	0.6585	0.4396
x2,x3,x4,x5,x6,***x*10**	0.9589	0.9195	0.8940	36.1754	2.08E-09	511.8	0.4813	0.4706
x2,***x*3**,x4,x5,x6	0.9577	0.9172	0.8965	44.3232	3.80E-10	**510.6**	0.5182	0.6497
x2,x4,***x*5**,x6	0.9533	0.9088	0.8914	52.3272	1.26E-10	511.1	0.4665	0.7968
x2,x4,x6	0.94950	0.9015	0.8881	67.1576	3.10E-11	511.1	0.5968	0.4702

Table 1.14: The dynamics of quality indicators during the process of explanatory variable selection in 2016

The model with the lowest AIC:

$$\begin{aligned}\hat{y}_t =& 7356.1238 - 0.2947 \cdot x_{2,t} - 3.3135 \cdot x_{3,t} + 1.2733 \cdot x_{4,t} \\ & - 0.0001 \cdot x_{5,t} + 0.00009029 \cdot x_{6,t}\end{aligned} \tag{1.7}$$

The removal of variables x_8, x_9, x_1, x_7 results in a model with the maximum value of R^2_{adj}, indicating their insignificant influence on the obtained indicator. However, testing the quality of coefficient estimates for individual variables using the Student's t-test does not allow us to consider the model highly significant, as the p-value for the t-statistic of b_3 is significantly stronger than the chosen α.

Upon excluding x_3, x_5, a regression model is obtained that includes only those factors whose coefficients are statistically significant at $\alpha = 0,05$. The model also exhibits satisfactory approximation quality (see Table 1.14) and takes the following form:

$$\underset{(t-Stat)}{\hat{y}_t \; =} \; \underset{(3.5024)}{9857.72} - \underset{(-4.2072)}{0.38741 \cdot x_{2,t}} - \underset{(8.3246)}{1.1236 \cdot x_{4,t}} + \underset{(2.9978)}{0.000059547 \cdot x_{6,t}} \tag{1.8}$$

The plots of the values y_i and $\hat{y}_i$ for the observation numbers $i, i = 1, \cdots, 26$ are displayed in Figure 1.2.

$E_{\text{rel.}} = 1,49\%$, hence, the approximation ability of the constructed model is relatively high, allowing for the practical reconstruction of the original data almost entirely.

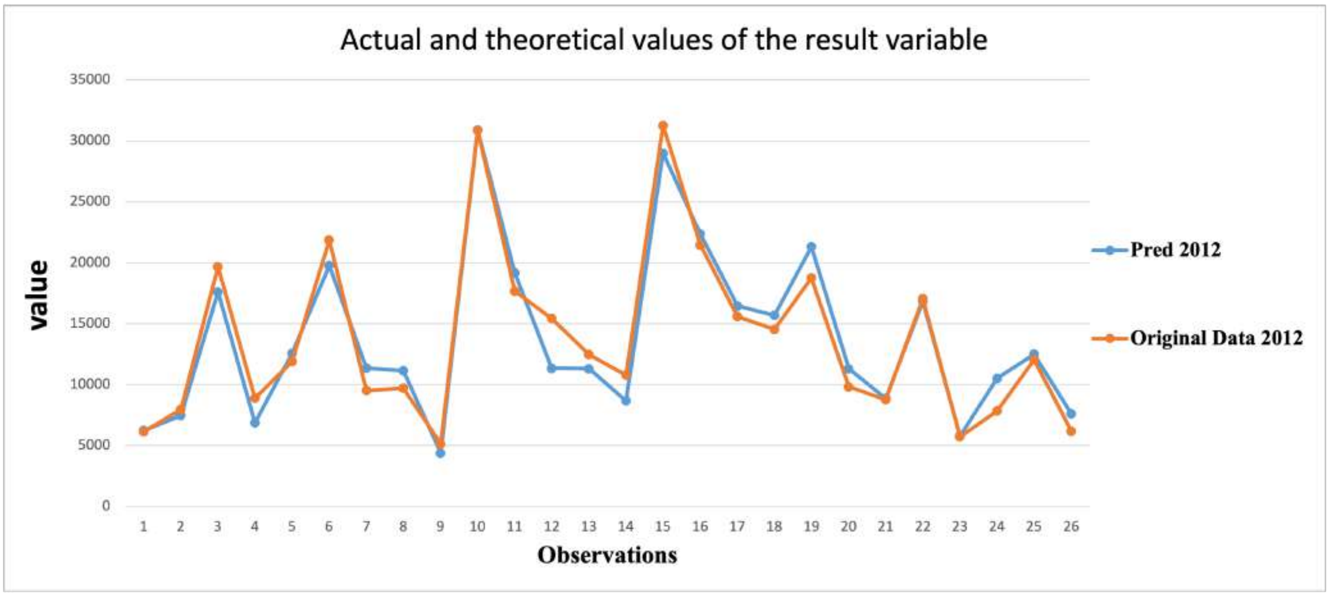


Figure 1.2

The relationship between the residuals of neighboring observations is weak, indicating the independence of the regression residuals. Therefore, the model represented in Figure 1.2, constructed based on the data from 2016, exhibits sufficient approximation capability. The quality estimates indicate the correctness of its specification, and the residual analysis confirms its adequacy.

The observed value of the Durbin-Watson statistic is $DW = 2,075$. For $\alpha = 0,05, n = 26$, and $k = 15$, the tabulated values are $d_L = 0,256$ and $d_U = 3,179$. Since $d_L < DW < d_U$, it is not possible to discern the presence or absence of autocorrelation.

Therefore, the model represented in Figure 1.2, constructed based on the data from 2016, exhibits sufficient approximation capability. The quality estimates indicate the correctness of its specification, and the residual analysis confirms its adequacy.

Multiple regression model for the years 2008-2015

In accordance with the aforementioned procedure employed to develop regression models for 2017, an analogous approach is applied to construct models individually for each year spanning from 2008 to 2015.

Multiple regression model for the year 2008

Regarding the specific year of 2008, after thorough analysis and processing of the available data, the resulting regression model is as follows:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.0111)}{1687.51} - \underset{(-2.7983)}{0.1026 \cdot x_{2,t}} - \underset{(24.2854)}{0.9663 \cdot x_{4,t}} \tag{1.9}$$

Hereafter (see Fig. 1.3), the plots illustrating the relationship between observed (y_i) and regression $(\hat{y}_i\ )$ values of the obtained indicator against the number of observations $i, i = 1, \cdots, 26$ are presented.

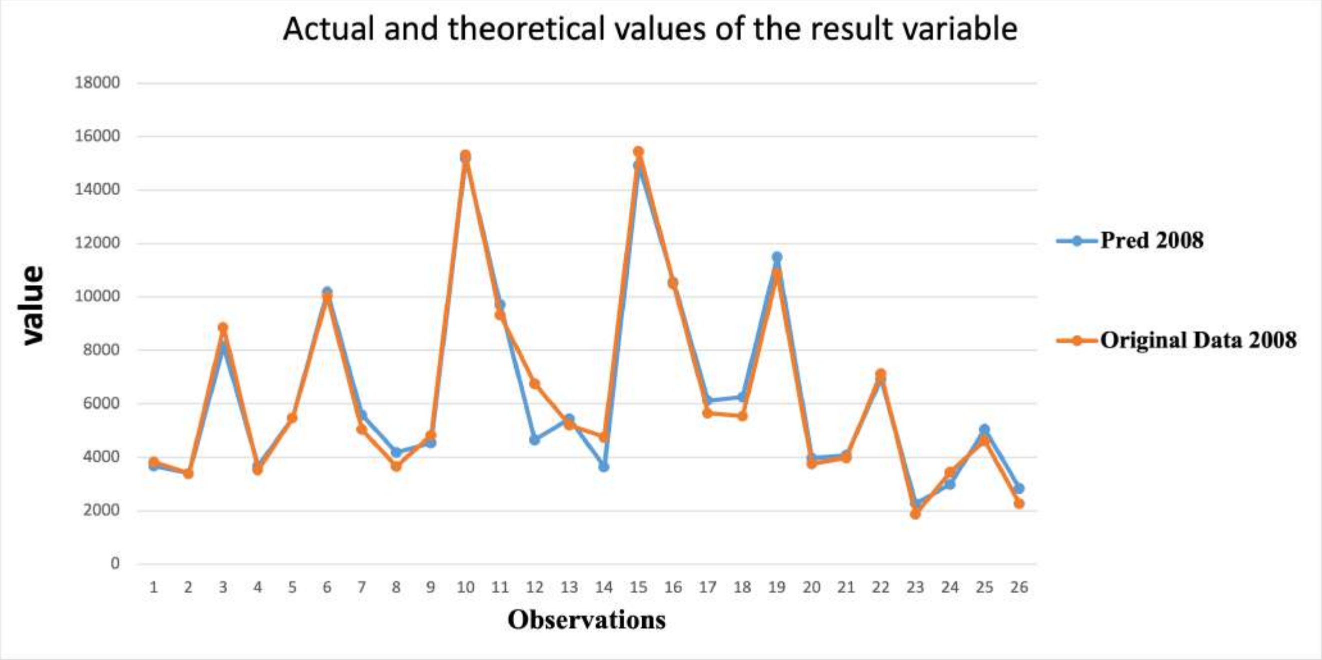


Figure 1.3

Table 1.15 presents the variations in the quality characteristics of the regression when selecting factors. The last row includes values of R^2,R,R^2_{adj}, and F for model (1.9).

With an approximate value close to 1, the magnitude of R^2_{adj} indicates a strong relationship between y and x_2. Furthermore, considering x_4 , we can also conclude that the model effectively approximates the observed values of the obtained variable, as indicated by $R = 0,9819$. The average relative approximation error is $E_{\text{rel.}} = 0,86\%$. The overall statistical significance of the regression model at a 5% level of significance is confirmed by the Fisher's test: the observed statistic $F = 309,1741$ significantly exceeds its critical value, $F_{(0,05;10;26-10-1)} = 2,543$.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9856	0.9715	0.9525	51.1401	9.45E-10	428.6	0.3739	0.2070
x1, x2, x3, x4, x5, x6, x8, x9, **x10**	0.9856	0.9714	0.9554	60.5808	1.16E-10	426.6	0.5827	0.2082
x1, x2, x3, x4, x5, x6, x8, **x9**	0.9856	0.9714	0.9581	72.3808	1.31E-11	424.6	0.5616	0.2093
x1, x2, x3, x4, x5, x6, x8	0.9855	0.9713	0.9601	87.0548	1.40E-12	422.8	0.5572	0.2486
x2, x3, x4,**x5**, x6, x8	0.9855	0.9712	0.9621	107.0218	1.29E-13	420.8	0.2571	0.2946
x2, **x3**, x4, x6, x8	0.9853	0.9708	0.9636	133.4335	1.19E-14	419.1	0.3136	0.2674
x2, x4, x6, **x8**	0.9839	0.9681	0.9620	159.3198	2.17E-15	419.5	0.2263	0.3083
x2, x4, **x6**	0.9831	0.9665	0.9620	212.0222	2.22E-16	418.7	0.3676	0.4269
x2, x4	0.9819	0.9641	0.9610	309.1741	2.39E-17	**418.6**	0.2075	0.3233

Table 1.15: The dynamics of quality indicators during the process of explanatory variable selection in 2008

Multiple regression model for the year 2009

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(4.2544)}{4642.4701} - \underset{(-3.9761)}{0.2680 \cdot x_{2,t}} + \underset{(11.1866)}{0.9244 \cdot x_{4,t}} + \underset{(2.3127)}{0.000039604 \cdot x_{6,t}} \tag{1.10}$$

The values of R^2,R,R^2_{adj}, and F for each model constructed through the sequential exclusion of statistically insignificant variables are presented in Table 1.16. Figure 1.4 depicts the differences in the dependent variable's values between the observed and calculated values using the regression model 1.10. The value of $E_{\text{rel.}} = 1,31\%$, indicating that the average approximation error of the studied data by the model is 1%.

Based on the values of R^2,R,R^2_{adj}, the quality of the constructed model (1.10) can be considered good. The statistical significance at a significance level of 5% was confirmed by the result of the F-test. Since $F = 102.59$(in Table 1.16), and $F_{(0,05;10;26-10-1)} = 2,543$, the inequality is satisfied:

$$F > F_{(0,05;10;26-10-1)}$$

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, **x3**, x4, x5, x6, x7, x8, x9, x10	0.9692	0.9394	0.8991	23.2695	2.39E-07	459.0	0.2118	0.2706
x1, x2, x4, x5, x6, **x7**, x8, x9, x10	0.9691	0.9393	0.9051	27.5183	4.42E-08	457.0	0.2020	0.2127
x1, x2, x4, x5, x6, **x8**, x9, x10	0.9691	0.9392	0.9107	32.8815	7.34E-09	455.0	0.2065	0.2755
x1, x2, x4, x5, x6, x9, **x10**	0.9687	0.9385	0.9146	39.2845	1.23E-09	453.4	0.2635	0.2137
x1, x2, x4, x5, x6, x9	0.9684	0.9379	0.9183	47.8724	1.82E-10	451.6	0.2518	0.2697
x2, x4, x5, x6,**x9**	0.9674	0.9359	0.9359	58.4359	3.02E-11	450.4	0.2692	0.3069
x2, x4, **x5**, x6	0.9668	0.9347	0.9222	75.1750	3.88E-12	448.9	0.2086	0.4119
x2, x4,x6	0.9661	0.9332	0.9241	102.5999	4.36E-13	**447.5**	0.2433	0.5307

Table 1.16: The dynamics of quality indicators during the process of explanatory variable selection in 2009

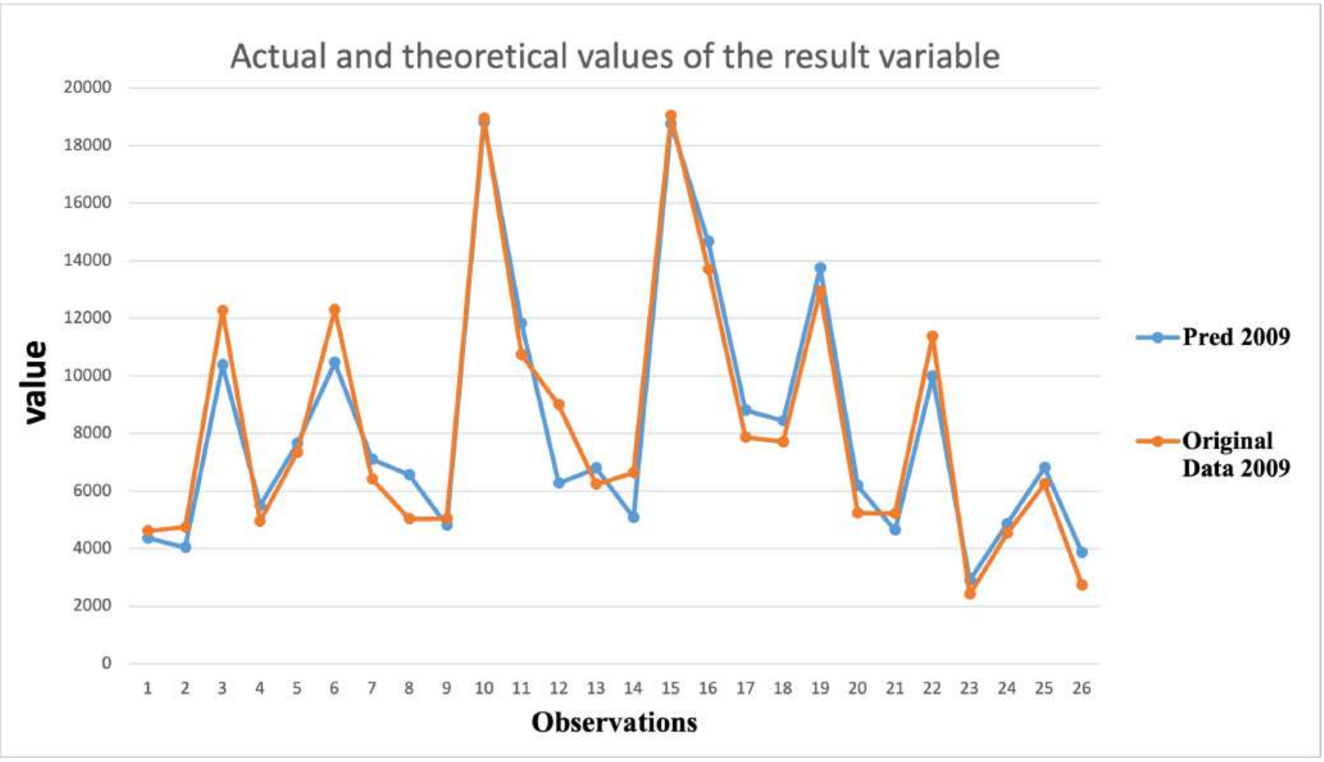


Figure 1.4

The observed value of the Durbin-Watson statistic is $DW = 1,9732$. For $\alpha = 0,05, n = 26, k = 15$, the tabulated values are $d_L = 0,256$ and $d_U = 3,179$. Since $d_L < DW < d_U$, it is not possible to determine the presence or absence of autocorrelation.

Multiple regression model for the year 2010

Now let's turn to the multiple regression model for the year 2010. After sequentially removing regressors, the following regression equation was obtained:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.2979)}{4682.4156} - \underset{(-2.9976)}{0.2274 \cdot x_{2,t}} + \underset{(14.3482)}{1.067 \cdot x_{4,t}} \tag{1.11}$$

The results of the regression are graphically depicted in Figure 1.5.

A comprehensive overview of the models constructed at each step of variable selection is presented in Table 1.17.

The model with the lowest AIC:

$$\hat{y}_t = 6030.6164 - 0.3115 \cdot x_{2,t} + 0.9447 \cdot x_{4,t} + 0.00003703 \cdot x_{6,t} \tag{1.12}$$

The qualitative characteristics for Table 1.17 are provided in the last row of the

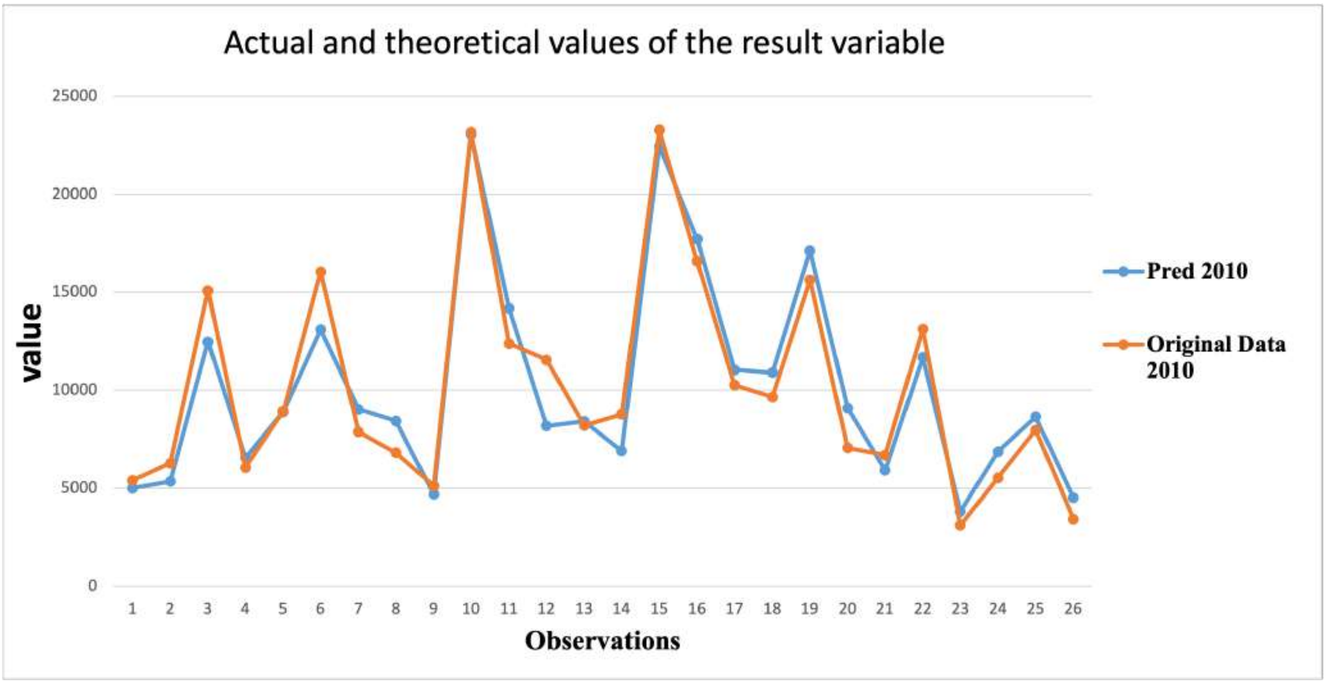


Figure 1.5

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, x7, **x8**, x9, x10	0.9609	0.9234	0.8723	18.0894	1.31×10^{-06}	475.0	0.2175	0.2643
x1, x2, x3, x4, x5, x6, x7, x9, **x10**	0.9608	0.9232	0.8801	21.3856	2.74×10^{-07}	473.1	0.1922	0.2715
x1, x2, x3, x4, x5, x6, **x7**, x9	0.9605	0.9226	0.8862	25.3526	5.48×10^{-08}	471.3	0.1979	0.2965
x1, x2, **x3**, x4, x5, x6, x9	0.9599	0.9214	0.8909	30.1653	1.07×10^{-08}	469.7	0.1805	0.3549
x1, x2, x4, **x5**, x6, x9	0.9594	0.9204	0.8953	36.6491	1.86×10^{-09}	468.0	0.2333	0.4098
x1, x2, x4, x6, **x9**	0.9584	0.9186	0.8982	45.1438	3.22×10^{-10}	466.6	0.2021	0.5352
x1, x2, x4, x6	0.9575	0.9169	0.9011	57.9834	4.76×10^{-11}	465.2	0.2187	0.6726
x2, x4, **x6**	0.9567	0.9153	0.9037	79.2873	5.94×10^{-12}	**463.7**	0.2266	0.4168
x2, x4	0.9489	0.9004	0.8917	103.9760	3.02×10^{-12}	465.9	0.3722	0.5357

Table 1.17: Dynamics of quality indicators during the process of explanatory variable selection in 2010

table. As evident, the values of the multiple correlation coefficient, coefficient of determination, and adjusted coefficient of determination are extremely high (>0.85), indicating a good quality of the constructed model.

The p-value for the observed F-statistic is $P - \text{value}(F) = 5.9492 \cdot 10^{-12}$, thus demonstrating the statistical significance of the model at $\alpha = 0.05$. The approximation error is negligible, with $E_{\text{rel.}} = 1,38\%$. To test the assumption of independence of regression residuals, the linear autocorrelation coefficient is computed.

Multiple regression model for the year 2011

For the year 2011, the following multiple regression model was obtained:

$$\underset{(t-Stat)}{\hat{y}_t =} \underset{(3.8588)}{6043.8067} - \underset{(-3.8596)}{0.2865 \cdot x_{2,t}} + \underset{(11.2310)}{0.8883 \cdot x_{4,t}} + \underset{(2.6593)}{0.000039970 \cdot x_{6,t}} \tag{1.13}$$

The average relative approximation error is $E_{\text{rel.}} = 1,16\%$. The model accurately reconstructs the observed values of y, as confirmed by the plots shown in Figure 1.6.

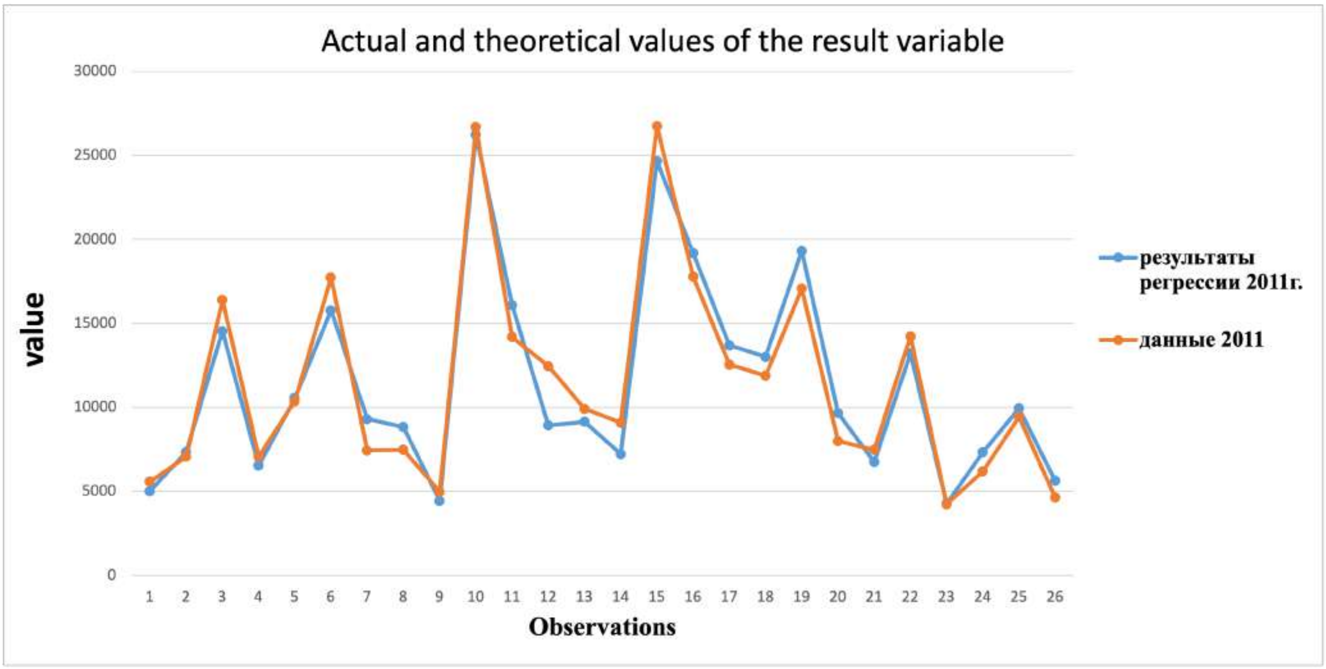


Figure 1.6

According to the data in Table 1.18, the final regression equation exhibits good quality estimates. The hypothesis that all coefficients of the model are equal to zero is rejected by the F-test. The observed value of the Fisher statistic significantly exceeds the tabulated value for this α ($F_{\text{obs}} = 102.2490, F_{(0,05;10;26-10-1)} = 2,543$), thus indicating the statistical significance of equation 1.13.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, **x3**, x4, x5, x6, x7, x8, x9, x10	0.9692	0.9394	0.8991	23.2695	2.39E-07	459.0	0.2118	0.2706
x1, x2, x4, x5, x6, **x7**, x8, x9, x10	0.9691	0.9393	0.9051	27.5183	4.42E-08	457.0	0.2020	0.2127
x1, x2, x4, x5, x6, **x8**, x9, x10	0.9691	0.9392	0.9107	32.8815	7.34E-09	455.0	0.2065	0.2755
x1, x2, x4, x5, x6, x9, **x10**	0.9687	0.9385	0.9146	39.2845	1.23E-09	453.4	0.2635	0.2137
x1, x2, x4, x5, x6, x9	0.9684	0.9379	0.9183	47.8724	1.82E-10	451.6	0.2518	0.2697
x2, x4, x5, x6, **x9**	0.9674	0.9359	0.9359	58.4359	3.02E-11	450.4	0.2692	0.3069
x2, x4, **x5**, x6	0.9668	0.9347	0.9222	75.1750	3.88E-12	448.9	0.2086	0.4119
x2, x4, x6	0.9661	0.9332	0.9241	102.5999	4.36E-13	**447.5**	0.2433	0.5307

Table 1.18: The dynamics of quality indicators during the process of explanatory variable selection in 2011

Multiple regression model for the year 2012

$$\underset{(t-Stat)}{\hat{y}_t =} \underset{(4.1382)}{8342.0697} - \underset{(-4.0912)}{0.34 \cdot x_{2,t}} + \underset{(10.0552)}{0.8786 \cdot x_{4,t}} + \underset{(3.5651)}{0.00005317 \cdot x_{6,t}} \tag{1.14}$$

Table 1.19 presents the values of quality criteria for the models constructed at each step of sequentially removing statistically insignificant parameters. The final model, in which all parameters are significant at the 5% level, is considered significant and possesses excellent estimates.

The multiple correlation coefficient is $R = 0.9606$, the coefficient of determination is $R^2 = 0.92287$, and with consideration for the number of factors, the adjusted coefficient of determination is $R^2_{adj} = 0.91235$. Thus, the model (1.14) explains approximately 90% of the variation in the observed values of the dependent variable $y_i (i = 1, \cdots, 26)$ The adequacy of the model at $\alpha = 0,05$ is confirmed by a low p-value of $P - \text{value}(F) = 2.1425 \cdot 10^{-12}$ (for the observed value of the Fisher statistic $F_{\text{obs}} = 87.7465$).

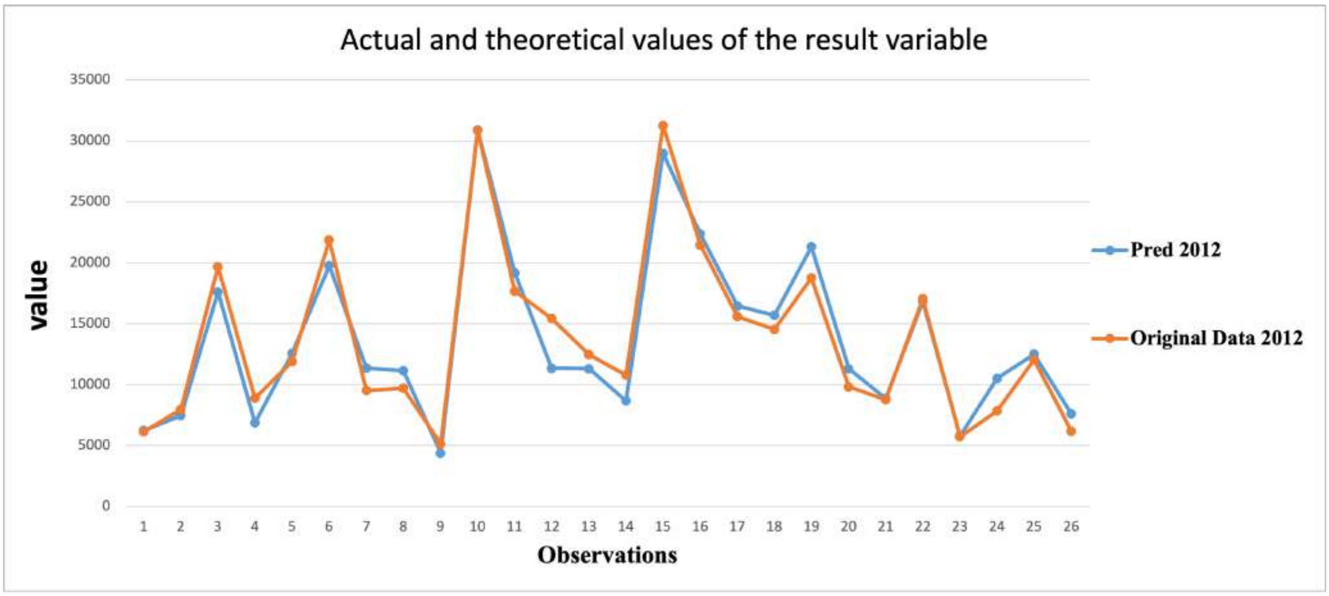


Figure 1.7

The plot depicting the relationship between the original values y_i and the number of observations $i (i = 1, \cdots, 26)$ and the plot showing the predicted values of model (1.14) as a function of $i (i = 1, \cdots, 26)$ are nearly convergent (see Figure 1.7). The obtained average relative error of 1.1% demonstrates high quality of approximation.

The model with the lowest AIC:

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9715	0.9438	0.9064	25.2225	1.38E-07	480.9	0.4570	0.6975
x1, x2, x3, x4, x5, x6, x8, x9, x10	0.9714	0.9437	0.9121	29.8452	2.43E-08	479.0	0.4082	0.7102
x2, x3, x4, x5, x6,x8, x9, **x10**	0.9713	0.9435	0.9169	35.5240	4.01E-09	477.1	0.4411	0.7444
x2, x3, x4, x5, x6,x8, **x9**	0.9711	0.94309	0.9209	42.6169	6.23E-10	475.3	0.4280	0.7197
x2, x3, x4, **x5**, x6,x8	0.9707	0.94235	0.9241	51.7688	9.15E-11	473.6	0.3685	0.5454
x2, **x3**, x4, x6,x8	0.9691	0.9391	0.9239	61.7467	1.81E-11	**473.0**	0.6762	0.8025
x2,x4, x6,**x8**	0.9623	0.9261	0.912	65.8038	1.41E-11	476.1	0.4179	0.7474
x2, x4, x6	0.9606	0.92287	0.91235	87.7465	2.14E-12	475.2	0.3052	0.6742

Table 1.19: The dynamics of quality indicators during the process of explanatory variable selection in 2012

$$\begin{aligned}\hat{y}_t =& 11830 - 0.4771 \cdot x_{2,t} - 6.157 \cdot x_{3,t} + 0.8173 \cdot x_{4,t} + 0.00006969 \cdot x_{6,t} \\ &+ 10.4689 \cdot x_{8,t}\end{aligned} \tag{1.15}$$

Multiple regression model for the year 2013

The multiple regression model for the year 2013, after sequentially excluding statistically insignificant parameters (based on Student's t-test with $\alpha = 0,05$), takes the following form:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{4.49705)}{10791.7171} - \underset{(-4.4098)}{0.3947 \cdot x_{2,t}} + \underset{(9.919)}{0.9138 \cdot x_{4,t}} + \underset{(4.0522)}{0.00005939 \cdot x_{6,t}} \tag{1.16}$$

The quality of the constructed models, corresponding to each step of sequential removal of insignificant factors, can be evaluated by analyzing the indicators in Table 1.20. The values of R , R^2 , and R^2_{adj} for the final model 1.16, presented in the last row, indicate its good quality. For $\alpha = 0,05$, the regression equation is considered statistically significant ($F_{\text{obs}} = 87.7465$,$F_{(0,05;10;26-10-1)} = 2,543$).

The approximation error of the model is negligibly small, with an average relative error of $E_{\text{rel.}} = 1,008\%$. The plots shown in Figure 1.8 confirm this.

The test for autocorrelation of residuals using the Durbin-Watson test does not provide a conclusive answer. Since $DW = 2,035$, for $\alpha = 0,05, n = 26, k = 15$, where $d_L = 0,256$ and $d_U = 3,179$ are the tabulated lower and upper critical values respectively, the observed statistic falls within the "inconclusive" range.

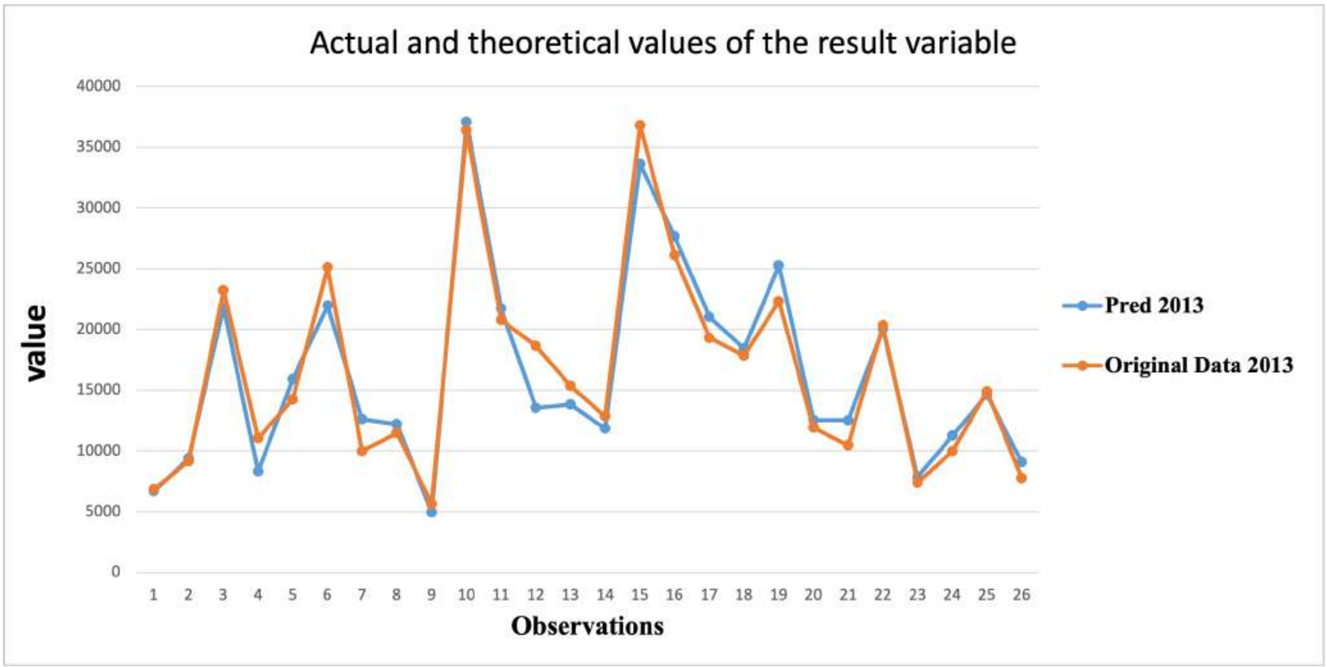


Figure 1.8

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9715	0.9438	0.9064	25.2225	1.38E-07	488.3	0.2052	0.2595
x1, x2, x3, x4, x5, x6, x8, x9, x10	0.9714	0.9437	0.9121	29.8452	2.43E-08	486.4	0.2015	0.3511
x2, x3, x4, x5, x6,x8, x9, **x10**	0.9713	0.9435	0.9169	35.5240	4.01E-09	485.4	0.2361	0.6711
x2, x3, x4, x5, x6,x8, **x9**	0.9711	0.94309	0.9209	42.6169	6.22E-10	484.6	0.3212	0.6277
x2, x3, x4, **x5**, x6,x8	0.9707	0.94235	0.9241	51.7688	9.15E-11	484.9	0.4562	0.3289
x2, **x3**, x4, x6,x8	0.9691	0.9391	0.9239	61.7467	1.81E-11	484.9	0.6160	0.7984
x2,x4, x6,**x8**	0.9623	0.9261	0.912	65.8038	1.41E-11	485.5	0.3986	0.2646
x2, x4, x6	0.9606	0.92287	0.91235	87.7465	2.14E-12	**484.2**	0.5032	0.3547

Table 1.20: The dynamics of quality indicators during the process of explanatory variable selection in 2013

Multiple regression model for the year 2014

The multiple regression model created using the data for the year 2014 contains the largest number of regressors and takes the following form:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(3.7025)}{7537.9438} - \underset{(-3.9412)}{0.3349 \cdot x_{2,t}} + \underset{(9.3550)}{0.9605 \cdot x_{4,t}} + \underset{(3.7005)}{0.00005521 \cdot x_{6,t}} \tag{1.17}$$

Table 1.21 presents the values of characteristics that allow evaluating the quality of each model in a general form, constructed based on the respective sets of explanatory factors formed by sequentially removing insignificant variables.

Model 1.17 possesses a relatively high adjusted coefficient of determination $R^2_{adj} = 0.904$, even with a large number of predictor variables. The multiple correlation coefficient is $R^2 = 0.9155$, indicating a strong relationship between the dependent variable and the set of regressors. The observed Fisher statistic corresponds to a p-value of $P - \text{value}(F) = 5.8 \cdot 10^{-12}$. The model is statistically significant at a 5% level of significance.

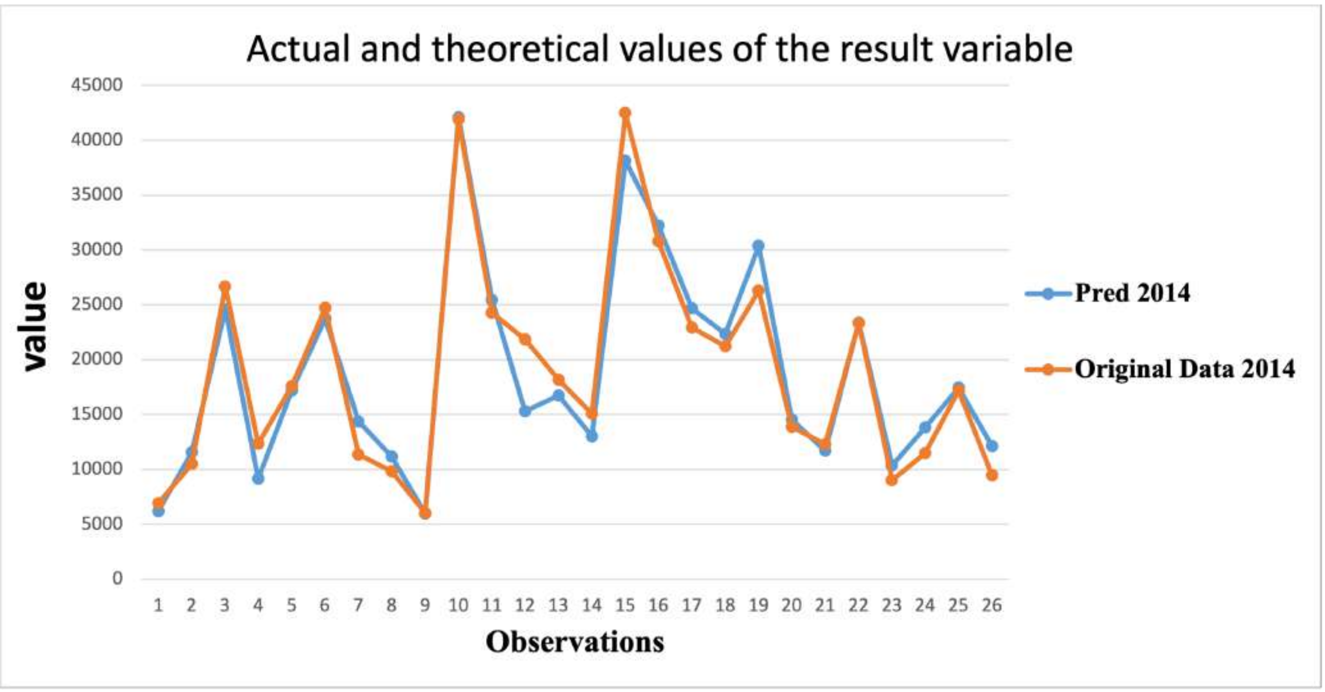


Figure 1.9

The satisfactory quality of the approximation of the initial data by the model is confirmed by the plots shown in Figure 1.9. The average relative error of approximation is $E_{\text{rel.}} = 1,06\%$.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9705	0.9419	0.9082	24.3349	1.76E-07	498.1	0.4011	0.3937
x2, x3, x4, **x5**, x6, x7, x8, x9, x10	0.9701	0.9412	0.9082	28.4854	3.43E-08	496.4	0.6827	0.3929
x2, x3, x4, x6, x7, **x8**, x9, x10	0.9687	0.9384	0.9095	32.4221	8.19E-09	494.7	0.7097	0.3543
x2, x3, x4, x6, x7, x9, **x10**	0.9686	0.9382	0.9142	39.0596	1.28E-09	**493.7**	0.2604	0.2153
x2, **x3**, x4, x6, x7, x9	0.9658	0.9329	0.9117	44.0432	3.79E-10	493.8	0.3784	0.5661
x2,x4, x6, x7, **x9**	0.9626	0.9267	0.9084	50.6182	1.13E-10	494.1	0.3487	0.5764
x2,x4, x6, **x7**	0.9589	0.9194	0.9041	59.9660	3.46E-11	494.6	0.6049	0.4713
x2, x4, x6	0.9568	0.9155	0.9040	79.4885	5.8E-12	493.8	0.9644	0.8195

Table 1.21: The dynamics of quality indicators during the process of explanatory variable selection in 2014

The model with the lowest AIC:

$$\begin{aligned} \hat{y}_t =& 12700 - 0.2101 \cdot x_{2,t} - 5.4055 \cdot x_{3,t} + 0.5423 \cdot x_{4,t} + 0.00009806 \cdot x_{6,t} \\ & - 0.0231 \cdot x_{7,t} + 0.5635 \cdot x_{9,t} - 1704.98 \cdot x_{10,t}. \end{aligned} \tag{1.18}$$

Multiple regression model for the year 2015

The construction of the model based on the 2015 dataset is considered. The selection of the most significant factors in the regression equation is presented in Table 1.22, which reflects the relationship between the values of quality characteristics of the regression and the set of exogenous variables.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9634	0.9281	0.8802	19.3829	8.27E-07	509.9	0.6264	0.2581
x1, x2, x3, x4, x5, x6, x8, x9, **x10**	0.9634	0.9281	0.88775	22.9691	1.64E-07	508.0	0.6017	0.2601
x1, x2, x3, x4, x5, x6, x8, x9	0.9633	0.9281	0.8942	27.4313	2.99E-08	506.0	0.6241	0.2720
x2, x3, x4, x5, x6, **x8**, x9	0.9632	0.9279	0.8998	33.1041	5.04E-09	504.0	0.6910	0.3098
x2, **x3**, x4, x5, x6, x9	0.9622	0.9259	0.9026	39.6144	9.54E-10	502.7	0.7131	0.3841
x2, x4, **x5**, x6,x9	0.96139	0.9242	0.9053	48.8209	1.58E-10	501.3	0.6327	0.3244
x2, x4, x6,**x9**	0.9598	0.9213	0.90635	61.4932	2.71E-11	500.3	0.8492	0.3608
x2, x4, x6	0.9595	0.9207	0.9099	85.2255	2.88E-12	**498.5**	0.8464	0.3921

Table 1.22: The dynamics of quality indicators during the process of explanatory variable selection in 2015

For the final stage, a model was obtained in which all parameters are statistically significant at the given level $\alpha = 0,05$, according to the significance of the factor estimates based on the Student's t-test. It takes the following form:

$$\hat{y_t} \underset{(t-Stat)}{=} \underset{(3.8623)}{8714.3617} - \underset{(-4.2869)}{0.3407 \cdot x_{2,t}} + \underset{(9.6562)}{1.0344 \cdot x_{4,t}} + \underset{(4.0281)}{0.00006283 \cdot x_{6,t}} \tag{1.19}$$

To assess the accuracy of approximation of the obtained regression model, a plot was constructed showing the initial values of the resulting indicator for the number of observations $i(i = 1, \cdots, 26)$, and another plot was generated displaying the values of the obtained indicator, calculated using the method 1.19, also for the number of observations $i(i = 1, \cdots, 26)$ (see Figure 1.10). It is evident that the plots are almost identical. The average relative error of approximation is $E_{\text{rel.}} = 1,22\%$, indicating that the error in the predicted values of lny is approximately 1%.

The adjusted coefficient of determination $R^2_{adj} = 0.9099$ also indicates the good approximation ability of the model and proves the existence of a functional relationship between the dependent variable (lny) and the explanatory variables (x_2, x_4, x_6). The magnitude of the multiple correlation coefficient is close to 1, indicating a very strong relationship. The adequacy of the model, or in other words, the hypothesis that all parameters are equal to zero, was tested using the F-test with the regressors. It was found that equation (1.19) is statistically significant in general at the given

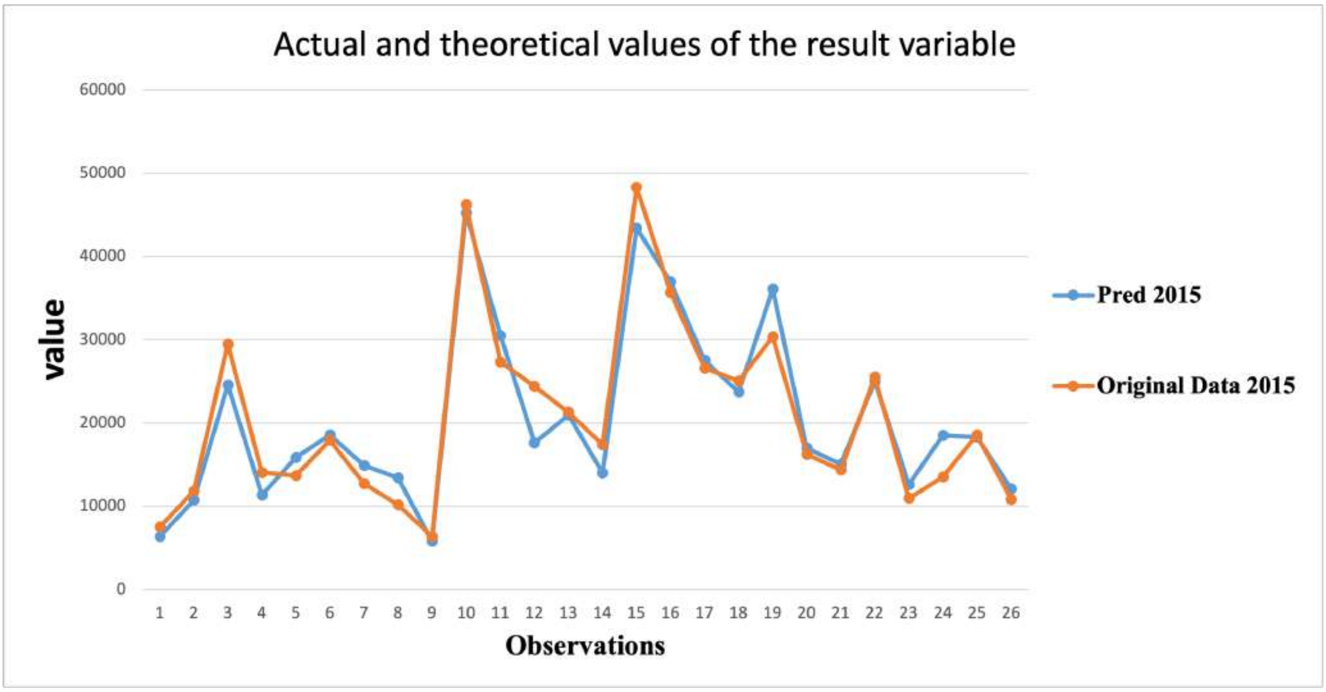


Figure 1.10

level $\alpha = 0,05(F = 85.2255)$. The observed value of the Durbin-Watson statistic is $DW = 1,814$, indicating the absence of residual autocorrelation. Thus, the multiple regression model (2.4) for the year 2009 is adequate, of good quality, and suitable for economic analysis.

1.2.6 Construction and Analysis of Models for the Investment Attractiveness of ASEAN-5 Regions

ASEAN preceded an organization established on July 31, 1961, called the Association of Southeast Asia (ASA), a group consisting of Thailand, the Philippines, and Malaysia.

Multiple regression models for Indonesia

Based on World Bank data for Indonesia from 1998 to 2014, a table is constructed to determine the parameters of the regression equation in the following format:

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), i = 1 \dots 17. \tag{1.20}$$

Where:i - the year index;
y_i- logarithm of the estimate of the volume of investments in year i;
x_{1i} - electricity consumption in year i;

country year	Indonesia	Malaysia	Singapore	Thailand	Philippines
1998	-0.241	0.163	7.314	7.315	2.287
1999	-1.866	3.895	16.578	6.103	1.247
2000	-4.55	3.788	15.515	3.366	1.487
2001	-2.977	0.5539	17.007	5.067	0.76
2002	0.1451	3.193	6.157	3.342	1.769
2003	-0.5969	3.219	17.051	5.232	0.492
2004	1.896	4.376	24.39	5.86	0.592
2005	8.336	3.925	18.09	8.216	1.664
2006	4.914	7.691	36.924	8.917	2.707
2007	6.928	9.071	47.733	8.634	2.919
2008	9.318	7.573	12.201	8.562	1.34
2009	4.877	0.1146	23.821	6.411	2.065
2010	15.292	10.886	55.076	14.747	1.07
2011	20.565	15.119	49.156	2.474	2.007
2012	21.201	8.896	55.31	12.899	3.215
2013	23.282	11.296	64.39	15.936	3.737
2014	25.121	10.619	68.698	4.975	5.74

Table 1.23: Investment attractiveness index (in trillions of dollars) for ASEAN-5.

x_{2i} - per capita income in year i;

x_{3i} - debt from loans provided to legal entities in year i;

x_{4i} - value of fixed assets in year i;

x_{5i} - volume of construction activity in year i;

x_{6i} - number of enterprises and organizations in year i;

x_{7i} - retail trade turnover in year i;

x_{8i} - GDP per capita in year i;

x_{9i} - unemployment rate in year i (in percentage).

The coefficient estimates are calculated using the regression analysis add-in tool in MS Excel for the observed values $y = (y_1, ..., y_{17})^\top$ and $X = (x_{1i}, x_{2i}, \cdots, x_{9i})^\top$, $i = 1, \cdots, 17$. The estimated coefficients take the form shown in Table 1.12:

The overall quality criteria of the model indicate a high level of its approximation ability. The coefficient of multiple determination is $R^2 = 0,9889$, while after adjustment for the penalty of having a significant number of explanatory variables, the adjusted coefficient of determination is $R^2_{adj} = 0,9746$. The multiple correlation

b0(c.o.)	208.8437	(90.64741)
b1(c.o.)	0.082245	(0.046266)
b2(c.o.)	-0.00285	(0.005001)
b3(c.o.)	0.103778	(0.10592)
b4(c.o.)	-0.28036	(0.497253)
b5(c.o.)	-2.84121	(2.320046)
b6(c.o.)	0.04912	(0.043409)
b7(c.o.)	0.012576	(0.093517)
b8(c.o.)	0.012724	(0.003133)
b9(c.o.)	-2.40566	(0.553048)

Table 1.24: The parameter estimates of the regression equation for Indonesia.

coefficient approaches unity with $R = 0,9944$ (see Table 1.25).

R	0,9944
R^2	0,9889
R^2_{adj}	0,9746

Table 1.25

The calculated value of the F-statistic is $F = 69,3286$. At a given significance level of $\alpha = 0,05$, the critical value is $F_{0,05;9;7} = 3,68$. Therefore, the model is statistically significant and practically explains most of the total variance in the outcome variable. However, despite the overall good quality of the regression model, the parameter estimates for the explanatory variables are not considered statistically significant at the chosen α level (see Table 1.26).

	t Stat	P-value
b0	2.303912	0.054675016
b1	1.777675	0.118697712
b2	-0.57053	0.586160124
b3	0.97978	0.359835289
b4	-0.56381	0.590481302
b5	-1.22463	0.260321937
b6	1.131562	0.295089445
b7	0.134478	0.896809381
b8	4.06166	0.004799517
b9	-4.34982	0.003355596

Table 1.26: The t-statistics for Indonesia.

Let us proceed with the methodology of stepwise exclusion of statistically insignificant factors from the regression equation. According to the obtained results, on the second stage, a model is built without x_7 ($b_7 = 0$). The new model has almost the same quality as the previous one, but it is statistically more significant according to the Fisher's F-test.

The dynamics of the changes in the indicators R^2,R, R^2_{adj} , and F during the sequential removal of insignificant factors from the multiple regression model are presented in Table 1.27.

The removal of variables x_7, x_2, x_4 leads to a model with the maximum value of R^2_{adj}, indicating their insignificant influence on the outcome variable. The assessment of coefficient quality for individual variables using the t-test does not allow

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9944	0.9888	0.9777	88.9045	5.21E-07	66.49	0.4568	0.3919
x1, x3, x4, x5, x6, x8, x9	0.9941	0.9883	0.9793	109.1355	5.61E-08	65.28	0.6183	0.4863
x1, x3, x5, x6, x8, x9	0.9935	0.9871	0.97936	127.5862	7.33E-09	**65.00**	0.6306	0.5709
x1, x5, x6, x8, x9	0.9924	0.9849	0.9781	144.4038	1.2E-09	65.49	0.4901	0.3775
x1, x5, x8, x9	0.9903	0.9807	0.9743	152.8212	3.51E-10	67.83	0.8697	0.4648

Table 1.27: The dynamics of quality indicators during the process of explanatory variable selection in Indonesia.

us to claim high significance of the model, as the p-value for the t-statistic of b_3 is significantly higher than the chosen $\alpha(0.2295 > 0.05)$.

After excluding x_3, x_6 , a regression model was obtained that includes only those factors whose coefficients are statistically significant at $\alpha = 0,05$. The model also demonstrates good approximation quality (see Table 1.27) and has the following form:

$$\begin{aligned}\underset{(t-Stat)}{\hat{y}_t} = \underset{(4.7593)}{8714.3617} + \underset{(6.1031)}{0.1190 \cdot x_{1,t}} - \underset{(-5.4129)}{4.7414 \cdot x_{5,t}} + \underset{(5.2747)}{0.01066 \cdot x_{8,t}} \\ - \underset{(-4.3683)}{2.1467 \cdot x_{9,t}}\end{aligned} \tag{1.21}$$

The plots of the values y_i and $\hat{y}_i$ for observations $i, i = 1, \cdots, 17$ are shown in Figure 1.2.

The relative error $E_{\text{rel.}} = 3,31\%$, indicating that the approximation ability of the constructed model is quite high, as it manages to almost fully recover the original data.

The calculated value of the Durbin-Watson statistic is $DW = 2,920$. For $\alpha = 0,05, n = 17, k = 7$, the critical values are $d_L = 0,451$ and $d_U = 2,537$. Since $4 - d_U < DW < 4 - d_L$, it is not possible to determine the presence or absence of autocorrelation.

The model with the lowest AIC:

$$\begin{aligned}\hat{y}_t =& 250.447 + 0.0828 \cdot x_{1,t} + 0.0499 \cdot x_{3,t} - 3.803 \cdot x5,t \\ &+ 0.048 \cdot x6,t + 0.0107 \cdot x_{8,t} - 2.4341 \cdot x_{9,t}\end{aligned} \tag{1.22}$$

Multiple regression models for Malaysia

According to the study conducted in Malaysia for 17 subjects (1998-2014), the following set of indicators was recorded, which characterizes the state and trends of

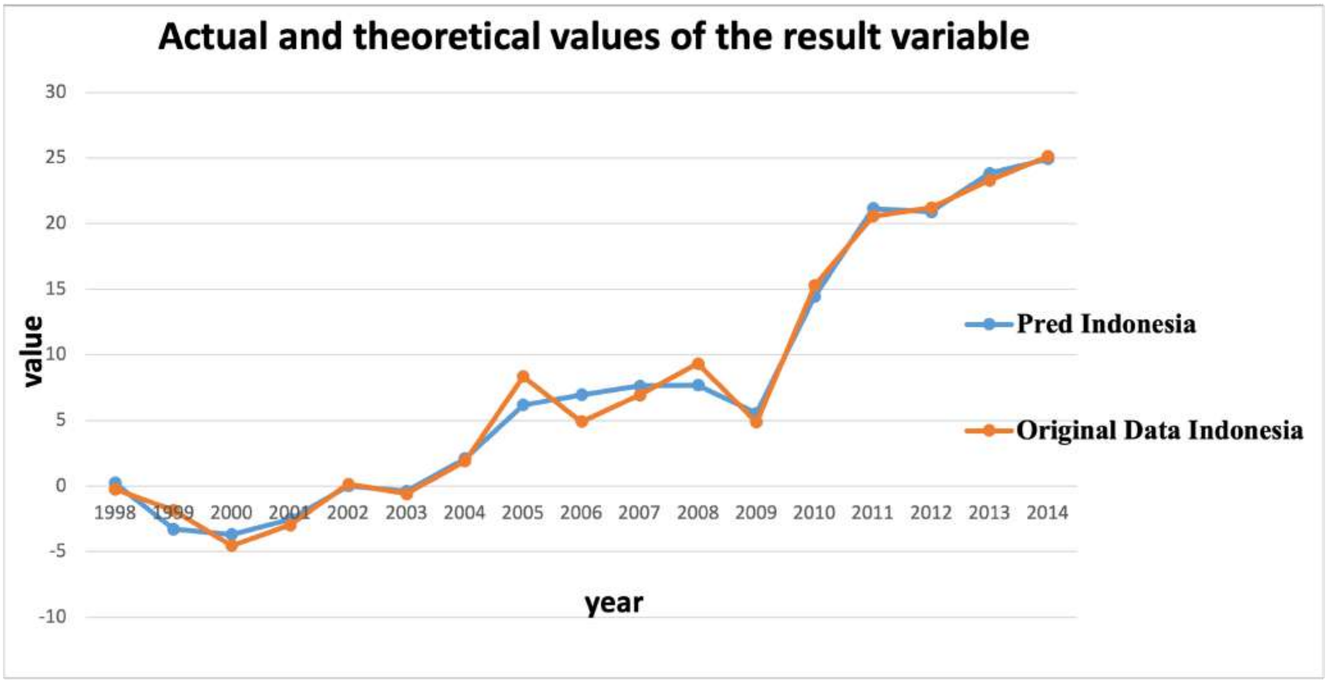


Figure 1.11

socio-economic development:

$$y_i = f(x_{1i}, x_{2i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), \quad i = 1..17. \tag{1.23}$$

Where:i - the year index;
y_i- logarithm of the estimate of the volume of investments in year i;
x_{1i} - electricity consumption in year i;
x_{2i} - per capita income in year i;
x_{3i} - debt from loans provided to legal entities in year i;
x_{4i} - value of fixed assets in year i;
x_{5i} - volume of construction activity in year i;
x_{6i} - number of enterprises and organizations in year i;
x_{7i} - retail trade turnover in year i;
x_{8i} - GDP per capita in year i;
x_{9i} - unemployment rate in year i (in percentage).

Let us discuss the construction of a model based on the Malaysia dataset. The selection of the most significant factors for the regression equation is presented in Table 1.28, which illustrates the dependency of the qualitative characteristics of the regression on the set of exogenous variables.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x4, x5, x6, **x7**, x8, x9	0.9574	0.9168	0.8336	11.0195	0.001364	72.06	0.3843	0.2903
x1, x2, x4, x5, x6, **x8**, x9	0.9572	0.9163	0.8512	14.0769	0.0003445	70.17	0.3948	0.2744
x1, x2, x5, x6, x8, **x9**	0.9557	0.9134	0.8615	17.5982	8.76E-05	**68.73**	0.2305	0.2689
x1, x2, x5, x6, x8	0.9333	0.8712	0.8127	14.885	0.0001407	73.49	0.3962	0.4037
x2, x5, **x6**, x8	0.9171	0.8411	0.7882	15.8888	9.706E-05	75.12	0.2522	0.2793
x2, **x5**, x8	0.9055	0.82	0.7784	19.74364	4.025E-05	75.28	0.3789	0.4228
x2, x8	0.8898	0.7917	0.762	26.6206	1.69E-05	75.77	0.4002	0.2107

Table 1.28: The dynamics of quality indicators during the process of explanatory variable selection in Malaysia.

In the final stage, a model was constructed in which all parameters are statistically significant at the chosen level of $\alpha = 0,05$, based on the significance of the factor estimates using the Student's t-test. It has the following form:

$$\hat{y}_t \underset{(t-Stat)}{=} \underset{(-1.4541)}{-1.9291} - \underset{(-2.8484)}{0.0038179 \cdot x_{2,t}} + \underset{(3.7713)}{0.0048033 \cdot x_{8,t}} \tag{1.24}$$

The adjusted coefficient of determination is $R^2_{adj} = 0.762$, which also indicates a good approximation ability of the model and suggests the existence of a functional relationship between the dependent variable (y) and the explanatory variables $(x_2, x_8$ $)$. The value of the multiple correlation coefficient is close to 1, indicating a strong correlation.

The adequacy of the model, or in other words, testing the hypothesis that all parameters are equal to zero using the F-test with the regressors, revealed that the equation 1.24 is statistically significant as a whole at the given $\alpha = 0,05$ level $(F_{\text{tab}} = 3,23,\ F_{\text{obs}} = 26.6206)$.

The observed value of the Durbin-Watson statistic is $DW = 2,414$. For $\alpha = 0,05, n = 17, k = 8,$, the critical values are $d_L = 0,356$ and $d_U = 2,757$. Since $d_L < DW < d_U$, it is not possible to determine the presence or absence of autocorrelation.

Therefore, the model 1.24, constructed based on the Malaysia dataset, exhibits sufficient approximation ability. The quality estimates indicate the correctness of the specification, and the residual analysis confirms its adequacy.

The model with the lowest AIC:

$$\begin{aligned}\hat{y}_t =& 75.9333 + 0.0115 \cdot x_{1,t} - 0.0051 \cdot x_{2,t} - 3.9452 \cdot x_{5,t} \\ &+ 0.066 \cdot x_{6,t} + 0.0079 \cdot x_{8,t} + 1.2757 \cdot x_{9,t}\end{aligned} \tag{1.25}$$

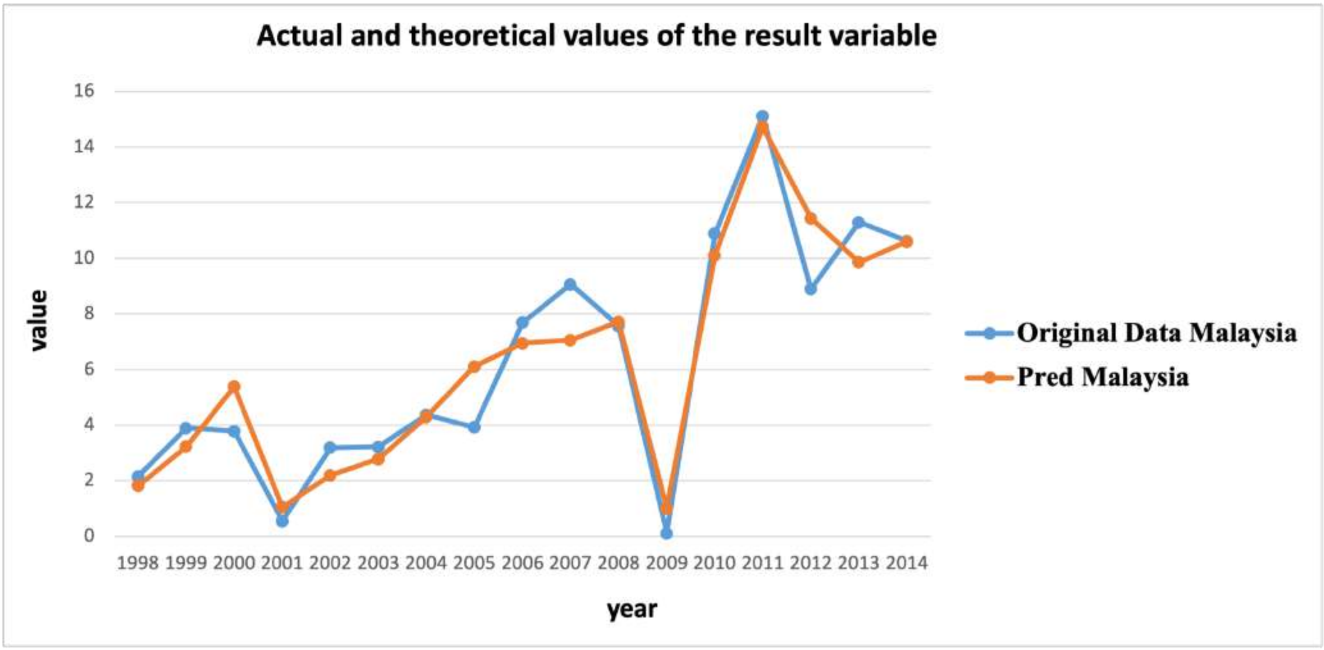


Figure 1.12

Multiple regression models for Singapore

Similar to the process described above for constructing a regression model for Indonesia and Malaysia, we will now construct a model for Singapore.

$$y_i = f(x_{1i}, x_{2i}, x_{4i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), \quad i = 1..17. \tag{1.26}$$

Where:i - the year index;

y_i- tlogarithm of the estimate of the volume of investments in year i;

x_{1i} - electricity consumption in year i;

x_{2i} - per capita income in year i;

x_{4i} - value of fixed assets in year i;

x_{6i} - number of enterprises and organizations in year i;

x_{7i} - retail trade turnover in year i;

x_{8i} - GDP per capita in year i;

x_{9i} - unemployment rate in year i (in percentage).

The final regression model obtained after data processing in Singapore is as follows:

$$\underset{(t-Stat)}{\hat{y}_t \;=} \underset{(-4.5665)}{-140.6215} + \underset{(8.6277)}{0.03392 \cdot x_{1,t}} - \underset{(-5.2671)}{0.372139 \cdot x_{7,t}} \tag{1.27}$$

The relationship between the observed (y_i) and predicted ($\hat{y}_i$) values of the outcome variable for each observed value $i, i = 1, \cdots, 17$, is shown below (see Figure 1.13).

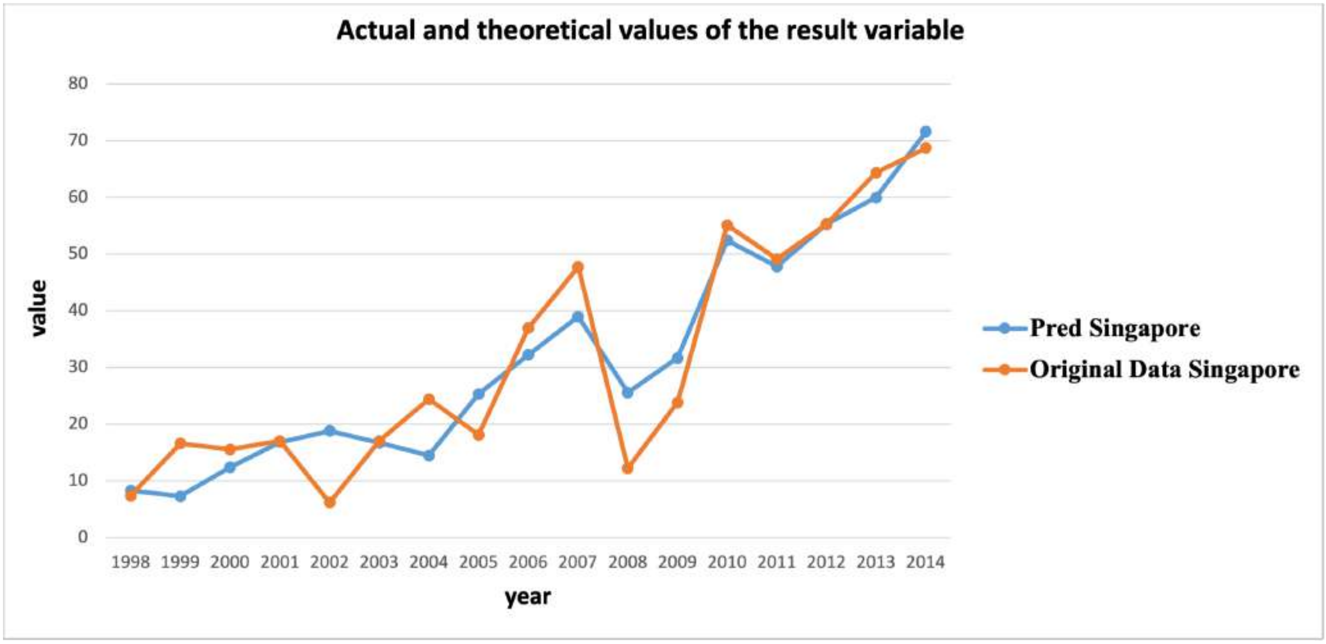


Figure 1.13: The change in qualitative characteristics of the regression as factors were selected

Table 1.29 illustrates the changes in the qualitative characteristics of the regression as factors were selected. The last row of the table includes the values of R, R^2, R^2_{adj}, and F_{obs} for the model in Table 1.29.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x4, x6, x7, x8, x9	0.9421	0.8876	0.8002	10.158	0.001216	129.4	0.5945	0.2602
x1, x2, x6, x7, x8, x9	0.942	0.8873	0.8198	13.132	0.0003125	127.4	0.5441	0.4628
x1, x2, x7, x8, x9	0.9411	0.8858	0.8339	17.0734	7.40E-05	125.7	0.6155	0.3796
x1, x7, x8, x9	0.9393	0.8824	0.8432	22.513	1.66E-05	124.2	0.7161	0.2220
x1, x7, x9	0.938	0.8799	0.8522	31.7596	2.98E-06	**122.5**	0.5762	0.2571
x1, x7	0.9214	0.849	0.8275	39.3794	1.78E-06	124.4	0.3717	0.2281

Table 1.29: The dynamics of quality indicators during the process of explanatory variable selection in Singapore.

The close-to-1 value of R^2_{adj} indicates a strong relationship between y and x_1, x_7. Furthermore, the model adequately approximates the observed values of the outcome variable, as evidenced by $R = 0.9214$. The overall statistical significance of the regression model at a 5% significance level is confirmed by the Fisher test: the observed statistic $F = 39.3794$ significantly exceeds its tabulated value of $F_{\text{tab}} = 3,29$.

The model with the lowest AIC:

$$\hat{y_t} = -309.2364 + 0.0277 \cdot x_{1,t} - 0.2632 \cdot x_{7,t} + 3.0053 \cdot x_{9,t} \tag{1.28}$$

Multiple regression models for Thailand

Following the approach described above for constructing a regression model in Indonesia, Malaysia, and Singapore, we will now proceed to construct a regression model in Thailand.

$$\begin{aligned} y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), \\ i = 1..17. \end{aligned} \tag{1.29}$$

Where:i - the year index;
y_i- logarithm of the estimate of the volume of investments in year i;
x_{1i} - electricity consumption in year i;
x_{2i} - per capita income in year i;
x_{3i} - debt from loans provided to legal entities in year i;
x_{4i} - value of fixed assets in year i;
x_{5i} - volume of construction activity in year i;
x_{6i} - number of enterprises and organizations in year i;
x_{7i} - retail trade turnover in year i;
x_{8i} - GDP per capita in year i;
x_{9i} - unemployment rate in year i (in percentage).

However, in the initial stage of regression analysis with all factors, the obtained R^2 and R^2_{adj} were very poor (far from 1) (see Table 1.30). Additionally, the p-value (F) was much larger than the significance level of $\alpha = 0,05$ (see Table 1.31).

R	0.599
R^2	0.5186
R^2_{adj}	0.474

Table 1.30

F-obs	F	p-value(F)
3.68	1.374583452	0.345016996

Table 1.31

These circumstances prevent us from constructing a regression model for Thailand.

Multiple regression models for Philippines

Based on the data provided by the World Bank for the Philippines, we will construct a computational table to determine the parameters of the regression equation in the following format:

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i}, x_{7i}, x_{8i}, x_{9i}), \qquad i = 1..17. \tag{1.30}$$

Where:i - the year index;
y_i- logarithm of the estimate of the volume of investments in year i;
x_{1i} - electricity consumption in year i;
x_{2i} - per capita income in year i;
x_{3i} - debt from loans provided to legal entities in year i;
x_{4i} - value of fixed assets in year i;
x_{5i} - volume of construction activity in year i;
x_{6i} - number of enterprises and organizations in year i;
x_{7i} - retail trade turnover in year i;
x_{8i} - GDP per capita in year i;
x_{9i} - unemployment rate in year i (in percentage).

The dynamics of the indicators R^2,R, R^2_{adj} and F when sequentially removing insignificant factors from the multiple regression model are presented in Table 1.32.

The removal of variables x_7 results in a model with the maximum value of $R^2_{adj} = 0.8678$, indicating their minor influence on the outcome variable. The evaluation of coefficient estimates for individual variables using the t-test does not confer high significance to the model, as the p-value for the t-statistic of b_4 is significantly higher than the chosen $\alpha(0.6027 > 0.05)$.

Explanatory variables	R	R^2	R^2_{adj}	F_{obs}	p-value(F)	AIC	p(RNT)	p(HT)
x1, x2, x3, x4, x5, x6, **x7**, x8, x9, x10	0.9665	0.9341	0.8495	11.0402	0.002268	30.13	0.2083	0.2672
x1, x2, x3, **x4**, x5, x6, x8, x9	0.9663	0.9339	0.8678	14.1291	0.0005677	28.34	0.2668	0.2935
x1, x2, x3, x5, x6, **x8**, x9	0.9602	0.9221	0.8615	15.2224	0.0002524	26.85	0.2879	0.3679
x1, x2, x3, x5, x6, **x9**	0.9319	0.8685	0.7896	11.0091	0.0006546	26.35	0.2819	0.3125
x1, x2, x3, x5, x6	0.9309	0.8666	0.8059	14.2942	0.0001696	25.77	0.2221	0.3256
x2, x3, x5, x6	0.9272	0.8597	0.8129	18.3863	4.69E-05	**25.34**	0.3795	0.2248

Table 1.32: The dynamics of quality indicators during the process of explanatory variable selection in Philippines.

After excluding x_4, x_8, x_9, a regression model was obtained that includes only those factors whose coefficients are statistically significant at $\alpha = 0,05$. The model also demonstrates good approximation quality (see Table 1.32) and takes the following form:

$$\begin{aligned}\underset{(t-Stat)}{\hat{y}_t} \; = \; & \underset{(-4.6988)}{-335.3628} - \underset{(-2.6266)}{0.003292 \cdot x_{2,t}} - \underset{(-3.9649)}{0.1683 \cdot x_{3,t}} + \underset{(4.9556)}{6.9177 \cdot x_{5,t}} \\ & + \underset{(2.20622)}{0.1421 \cdot x_{6,t}}\end{aligned} \tag{1.31}$$

The graphs of the observed values y_i and predicted values $\hat{y}_i$ for the observation numbers $i, i = 1, \cdots, 17$ are shown in Figure 1.14.

The relative error $E_{\text{rel.}} = 9,44\%$, indicating that the approximation ability of the constructed model is quite high, as it successfully recovers almost all of the original data.

The calculated value of the Durbin-Watson statistic is $DW = 2,473$. For $\alpha = 0,05, n = 17, k = 7$, the critical values are $d_L = 0,451$ and $d_U = 2,537$. Since $d_L < DW < d_U$, it is not possible to determine the presence or absence of autocorrelation.

Therefore, the model 1.31, constructed using data from Indonesia, exhibits sufficient approximation capability. The quality estimates indicate the correctness of the specification, and the residual analysis confirms its adequacy.

1.3 Discussion and Analysis

1.3.1 Discussion and Analysis of the Results for China

The data obtained from the Chinese statistical yearbook take into account the inflation level (all values are recorded at comparable prices).

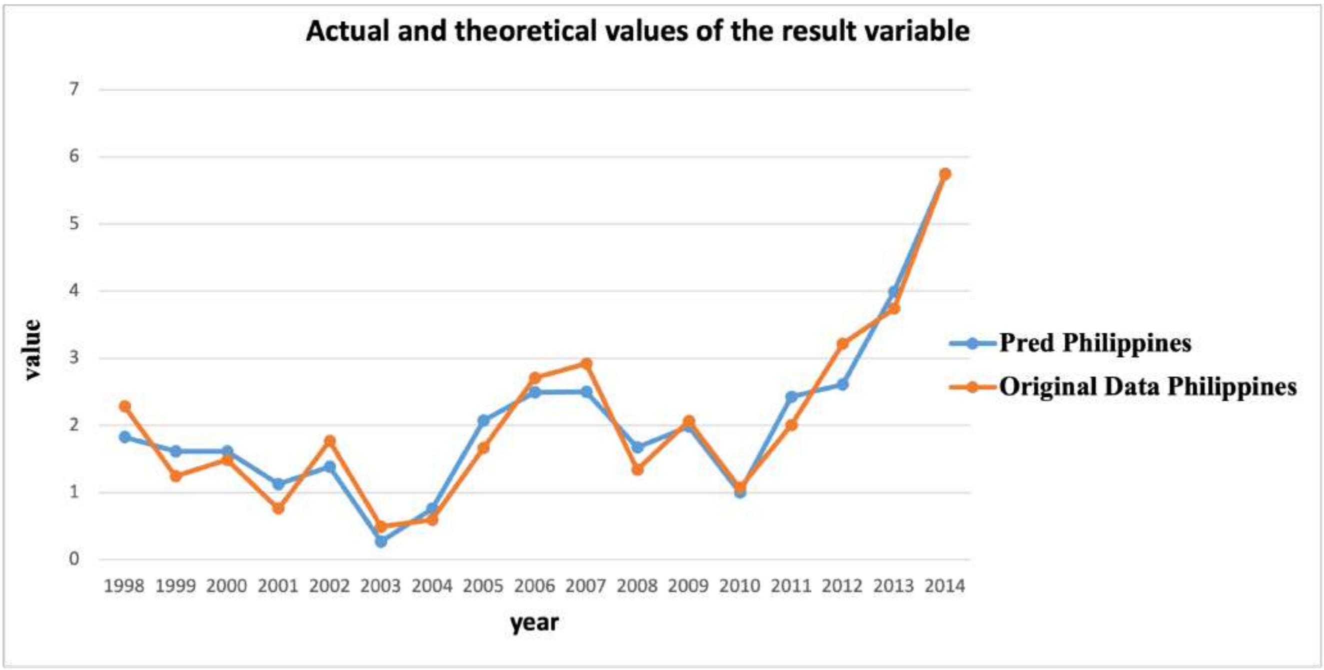
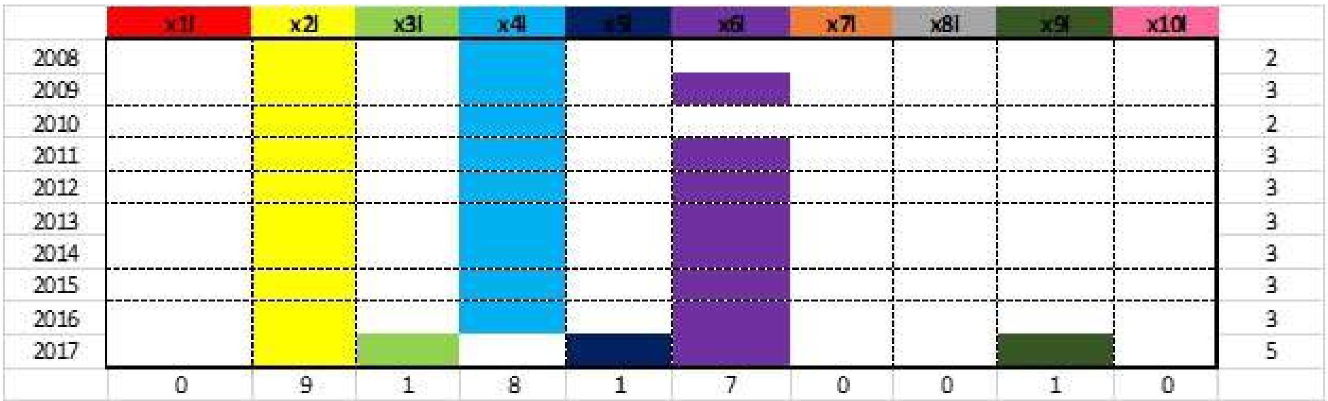


Figure 1.14

The interpretation of the results obtained is a crucial aspect of applied statistical research. After processing the collected data, the mathematical model that relates annual investments in the region to various factors representing different aspects of its socio-economic potential should be interpreted in economic terms.

In the statistical analysis of China's investment attractiveness, a multivariate regression model was constructed based on the stepwise elimination method using data from 2008 to 2017. Ten regression equations were developed for the 26 regions in China with the highest investments in 2018. By comparing the changes in socio-economic indicators over the past decade and considering their cumulative effect, significant impacts on the average cost of investments can be observed. The results are presented below (see Figure 1.15):



	x1l	x2l	x3l	x4l	x5l	x6l	x7l	x8l	x9l	x10l	
2008		■		■							2
2009		■		■		■					3
2010		■		■							2
2011		■		■		■					3
2012		■		■		■					3
2013		■		■		■					3
2014		■		■		■					3
2015		■		■		■					3
2016		■		■		■					3
2017		■	■		■	■			■		5
	0	9	1	8	1	7	0	0	1	0	

Figure 1.15

The factors correspond to the columns of the table (for clarity, each factor is assigned a specific cell color), while the rows represent sets of exogenous variables in the final regression model for each respective year. At the bottom of the table, the number of inclusions of the k-th factor ($k = 1, ..., 10$) in the regressors of the final models constructed for the years 2008-2017 is indicated. The number to the right of the table represents the count of explanatory variables in the final model for the corresponding year.

Regarding the factor x_2, it is included in all models, indicating its continuous influence on the formation of regional investments from 2008 to 2017. However, factors x_1, x_7, x_8, x_{10} are never included in the regression equation, suggesting that these indicators have minimal impact on investment volume during the period from 2008 to 2017.

The factor x_2 - per capita income - is included in all models. This indicates that per capita income is an important factor influencing investment attractiveness and has a positive impact on investment appeal.

The next most frequently included variable is x_4 - the value of fixed assets. It is logical that the value of fixed assets is one of the important factors influencing a country's investment attractiveness and plays a significant role in its overall socio-economic framework. However, in 2017, the growth of investments slowed down, increasing by 7.5% during the first three quarters and decreasing by 0.7 percentage points on an annual basis. Investment operations demonstrate new signs of "stability," with positive changes in several indicators such as funding sources, investments in the civil industry, and investments in equipment production. Currently, the economic benefits from investments have significantly decreased as budget consolidation curtails government and infrastructure investments, financial expenditure reductions restrict financing, the real estate market has slowed down, some sector-specific investments continue to decline, and other issues become more prominent.

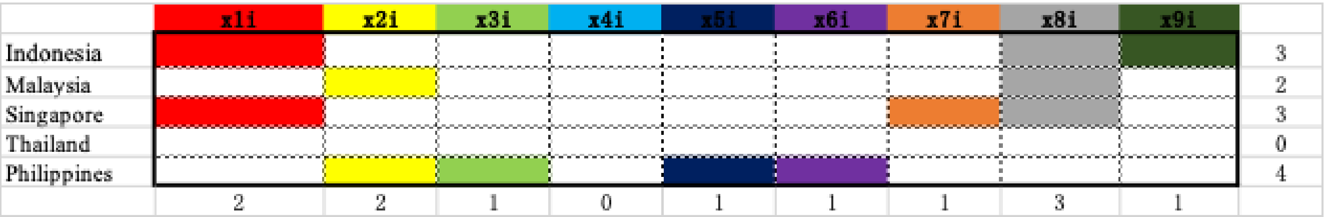
The factor x_6 - the volume of construction activities - also holds importance in the regression variables. It is highly likely that the examined regions experienced growth in the construction market during the studied period. As evident, many consider real estate investments to be among the most reliable. The absence of factor x_6 in the model for 2008 and 2010 may be attributed to the peak point of the 2008 crisis and high economic risks.

Overall, these findings lead to the conclusion that the unstable global economic situation significantly influences the formation of investment prerequisites at the micro and median levels of the country. Equally important is the current stage of the industrial cycle. Economic risk indicators are particularly crucial in the pre-crisis and post-crisis periods. However, under conditions of stable exports, substantial growth in industrial profits, active investment demand for scientific and technological reforms, and accelerated development of new individual investments, the volume of attracted investments should continue to increase.

1.3.2 Discussion and Analysis of the Results for ASEAN-5

Applied statistical research aims to interpret the results obtained. After data processing, it is necessary to explain, from an economic perspective, the mathematical model that describes the relationship between a country's investments and various factors representing different aspects of its socio-economic potential.

Based on the statistical analysis of investment attractiveness in ASEAN, a multivariate regression model was developed using the stepwise elimination method with data from 1998-2014. Five regression equations were constructed for the five ASEAN countries. Over the past 18 years, many socio-economic indicators have changed, and their cumulative effect has significantly influenced the average cost of investments.



	x1i	x2i	x3i	x4i	x5i	x6i	x7i	x8i	x9i	
Indonesia	🟥							▨	🟩	3
Malaysia		🟨						▨		2
Singapore	🟥						🟧	▨		3
Thailand										0
Philippines		🟨	🟩		🟦	🟪				4
	2	2	1	0	1	1	1	3	1	

Figure 1.16

For the five ASEAN countries, each factor has different significance due to variations in economic systems and national policies. However, x_8 - GDP per capita - is an important factor for ASEAN as a whole.

As for Thailand, in recent years, it has attracted a significant number of foreign tourists, and the volume of foreign investments has been growing annually. Currently, Thailand serves as an investment hub within ASEAN, with its policy primarily focused on investments in infrastructure and the digital economy. Therefore, it may not fit into this particular model.

ASEAN countries affected by the current global economic downturn are accelerating their internal economic integration process to address the challenges of slowing economic growth. It is necessary to strengthen economic exchanges and cooperation with the world and rely more on global investment flows. The investment market potential of ASEAN will continue to expand. Investments in ASEAN help countries utilize the region's abundant energy, labor, and other unique resources, contributing to the improvement and modernization of internal industrial structures and sustainable economic development.

The level of foreign direct investments (FDI) inflows in ASEAN countries positively correlates with factors such as the size of their economies, levels of natural resources and technology. Negative correlations can be observed with internal conflicts within ASEAN countries, trade freedom, investment freedom, financial freedom, and the impact of bilateral investment agreements on investment efficiency. However, the influence of government stability in ASEAN and infrastructure improvements can significantly enhance investment efficiency. Economic development level, research and development level, foreign capital inflows, and foreign investments have a positive impact, while monopolistic financial systems and trade barriers exert a negative influence on FDI. Taxation also plays a significant role in determining the level of foreign direct investments.

1.4 Conclusion to Chapter 1

In conclusion, this chapter utilizes multiple linear regression analysis to investigate the investment attractiveness of China and the ASEAN-5 countries. Through a stepwise regression model, the most influential factors affecting investment attractiveness are identified for each country.

The results reveal that per capita income (x_2) consistently influences regional investments from 2008 to 2017, indicating its continuous positive impact on investment appeal. On the other hand, factors such as population size (x_1), government stability (x_7), GDP per capita (x_8), and foreign exchange reserves (x_4) show minimal impact on investment volume during the study period.

Another important factor found through the analysis is the value of fixed assets

(x_4), which significantly influences a country's investment attractiveness and plays a vital role within its socio-economic framework. However, in 2017, the growth of investments slowed down due to various challenges, including budget consolidation, reduced financial expenditure, a slowdown in the real estate market, and declining sector-specific investments.

The volume of construction activities (x_6) also emerges as a significant variable in the regression models, indicating the importance of growth in the construction sector for investment attractiveness. The absence of x_6 in the models for 2008 and 2010 can be attributed to the economic risks associated with the global financial crisis during those years.

Overall, the findings highlight the substantial influence of the unstable global economic situation on investment prerequisites at both the micro and median levels of a country. The current stage of the industrial cycle and economic risk indicators play crucial roles, particularly during pre-crisis and post-crisis periods. However, under conditions of stable exports, substantial growth in industrial profits, active investment demand for scientific and technological reforms, and accelerated development of new individual investments, it is expected that the volume of attracted investments will continue to increase.

For the ASEAN-5 countries, each factor holds different significance due to variations in economic systems and national policies. However, GDP per capita (x_8) emerges as an important factor for ASEAN as a whole.

Regarding Thailand, its recent growth in foreign tourists and increasing foreign investments position it as an investment hub within ASEAN. It focuses primarily on investments in infrastructure and the digital economy. Consequently, Thailand may not fit into the specific model used in this study.

The ASEAN countries, affected by the current global economic downturn, are actively accelerating their internal economic integration process to address the challenges of slowing economic growth. Strengthening economic exchanges and cooperation with the world, as well as relying more on global investment flows, is deemed necessary. The potential of the ASEAN investment market is expected to expand, allowing countries to leverage the region's abundant energy, labor, and unique resources to improve and modernize internal industrial structures, leading to sustainable economic development.

Furthermore, the analysis reveals that foreign direct investment (FDI) inflows in ASEAN countries positively correlate with factors such as the size of their economies, levels of natural resources and technology. Conversely, negative correlations are observed with factors like internal conflicts, trade freedom, investment freedom, financial freedom, and the impact of bilateral investment agreements on investment efficiency. Government stability within ASEAN and infrastructure improvements can significantly enhance investment efficiency. Additionally, economic development level, research and development capability, foreign capital inflows, foreign investments, and favorable taxation policies have a positive impact on FDI. On the other hand, monopolistic financial systems and trade barriers exert a negative influence on FDI.

In summary, these findings provide valuable insights for policymakers and investors, guiding them in making informed decisions and formulating effective strategies to promote economic growth and development in China and the ASEAN-5 countries.

Enhancing Investment Attractiveness in China and ASEAN-5: Academic Perspectives and Policy Recommendations for Sustainable Economic Development

Based on the analysis and findings presented in this chapter, several suggestions can be made to further enhance investment attractiveness in China and the ASEAN-5 countries:

Strengthen policy reforms: Policymakers should focus on implementing comprehensive reforms to improve the investment environment. This includes streamlining regulations, reducing bureaucracy, enhancing transparency, and ensuring fair competition. Clear and enforceable rules and regulations will provide investors with confidence and encourage more foreign direct investment.

Promote economic diversification: Countries should strive for economic diversification to reduce reliance on a single sector or industry. This allows for a more stable and resilient economy, attracting a broader range of investments. Governments should identify and support emerging industries with high growth potential through targeted policies, incentives, and infrastructure development.

Enhance infrastructure development: Improving infrastructure is essential for attracting investments. Governments should prioritize infrastructure projects that facilitate trade, connectivity, and logistics. Developing reliable transportation networks, modernizing ports, and expanding telecommunications infrastructure will create an enabling environment for businesses and promote investment opportunities.

Foster research and development capabilities: Investing in programs and innovation is crucial for attracting high-value investments. Governments should allocate resources to develop indigenous technology and encourage collaboration between academia, industry, and research institutions. Providing incentives for companies to invest in activities will stimulate technological advancements and attract knowledge-based investments.

Enhance regional integration: Deepening economic integration within ASEAN and strengthening cooperation with external partners can create a more attractive investment environment. Harmonizing regulations, reducing trade barriers, and promoting regional economic cooperation frameworks will expand market access and increase investment flows within the region. Engaging in regional trade agreements and actively participating in global value chains will boost competitiveness and attract foreign investments.

Improve financial systems: Creating a robust and transparent financial system is essential for investment attractiveness. Governments should work towards developing efficient banking and capital markets, ensuring access to affordable financing for businesses. Promoting financial market liberalization, strengthening investor protection measures, and enhancing corporate governance practices will instill confidence in investors and encourage long-term investments.

Invest in human capital: Developing a skilled and educated workforce is crucial for attracting investments that require specialized knowledge and expertise. Governments should invest in education and vocational training programs to meet the demands of industries. Additionally, promoting entrepreneurship and fostering a culture of innovation will create a favorable environment for investment in high-growth sectors.

Improve ease of doing business: Governments should continue efforts to improve the ease of doing business by simplifying administrative procedures, reducing bureaucratic hurdles, and providing efficient public services. Establishing dedicated investment promotion agencies can offer one-stop services for investors, providing assistance throughout the investment process and resolving any challenges they may encounter.

Strengthen regional collaboration: China and the ASEAN-5 countries should collaborate closely on investment promotion initiatives. Sharing best practices, exchanging information, and facilitating cross-border investments will enhance the overall investment climate in the region. Joint marketing efforts, investment forums, and business matchmaking events can attract potential investors and foster economic cooperation.

Environmental sustainability: Emphasizing environmental sustainability and promoting green investments can enhance the attractiveness of a country or region. Implementing policies that prioritize renewable energy, sustainable development, and environmentally friendly practices will appeal to socially responsible investors and contribute to long-term economic growth.

By implementing these suggestions, policymakers can create a more conducive investment environment, attract a diverse range of investments, and foster sustainable economic development in China and the ASEAN-5 countries. It is vital to continuously evaluate and adapt policies to meet the changing needs and dynamics of the global investment landscape.

Chapter 2

Exploring Investment Attractiveness: An In-depth Analysis Using Cluster Analysis Method

This chapter provides a comprehensive overview of the principles, assumptions, and application scenarios of cluster analysis. Additionally, it demonstrates the application of cluster analysis using the R language to study the investment attractiveness of different regions in China. The analysis involves generating heat maps for visualizing the results and includes the corresponding algorithms. The data collection and preprocessing procedures used are consistent with those described in Chapter 1 Section 2.

Results presented in the chapter are published in paper [99]

2.1 Data and Methods

2.1.1 Methodology

Cluster analysis is an extensively employed method of unsupervised learning that aims to partition comparable observations into mutually exclusive clusters. By recognizing inherent patterns and structures in the data, it facilitates exploratory analysis and unveils concealed relationships or features.

The primary objective of cluster analysis is to determine the similarity between samples by utilizing similarity or distance measures, subsequently allocating samples to distinct clusters based on the degree of similarity. This process encompasses two vital steps: distance measurement and clustering algorithm.

Distance measurement involves assessing the likeness or dissimilarity between samples through the utilization of distance metrics such as Euclidean distance, Manhattan distance, and Minkowski distance. The selection of appropriate metrics depends on the specific requirements of the problem at hand.

Here are some commonly used mathematical formulas and metrics for calculating the similarity or distance between data points:

Euclidean Distance: Euclidean distance is the most commonly used distance metric, which calculates the straight-line distance between two N-dimensional data points. For two data points $x = (x_1, x_2, \cdots, x_N)$ and $y = (y_1, y_2, \cdots, y_N)$, the Euclidean distance can be represented as:

$$d_{euclidean}(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}$$

Manhattan Distance: Manhattan distance is another common distance metric, which calculates the straight-line distance between two N-dimensional data points but only considers vertical and horizontal movements on the coordinate axes. For two data points $x = (x_1, x_2, \cdots, x_N)$ and $y = (y_1, y_2, \cdots, y_N)$, the Manhattan distance can be represented as:

$$d_{manhattan}(x, y) = \sum_{i=1}^{N} |x_i - y_i|$$

Minkowski Distance: Minkowski distance is a generalized form of Euclidean distance and Manhattan distance. For two data points$x = (x_1, x_2, \cdots, x_N)$ and $y = (y_1, y_2, \cdots, y_N)$, the Minkowski distance can be represented as:

$$d_{minkowski}(x, y) = \left(\sum_{i=1}^{N} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

where the parameter p determines the shape of the distance metric. When $p = 1$, the Minkowski distance is equivalent to the Manhattan distance, and when $p = 2$, it is equivalent to the Euclidean distance.

These distance metrics can be used in clustering algorithms such as K-means clustering, hierarchical clustering, and density-based clustering. By selecting an appropriate distance metric and clustering algorithm, we can group data points into clusters based on their similarity or distance, thereby revealing the underlying structure and patterns within a dataset.

Clustering algorithms, on the other hand, establish distinct clusters based on the calculated distances. Commonly employed algorithms include K-means clustering, hierarchical clustering, and density clustering. K-means clustering is suited for situations where the number of clusters is predetermined, whereas hierarchical clustering generates clusters with a hierarchical structure.

Cluster analysis serves as a powerful technique for data analysis by identifying underlying patterns and structures, enabling the categorization of similar observations into distinct clusters. Familiarity with the principles and methods of cluster analysis enhances the ability to extract valuable insights from data, offering guidance in problem-solving endeavors.

2.2 Experiments and Results

2.2.1 Foundation for Model Construction

In order to ensure the validity of the results, it is imperative to establish certain criteria for the econometric model employed in this statistical study, emphasizing the need for quantifiable factors. The primary objective of this investigation is to develop models capable of assessing the effectiveness of investments in different regions, taking into consideration geographical disparities. It is assumed that investment activity is influenced by certain investment conditions, thus suggesting the utilization of an indicator reflecting the volume of fixed capital investment in the region as a dependent variable.

It is crucial to acknowledge that the volume of fixed capital investment encompasses various economic dimensions impacted by a wide range of socio-economic characteristics. Hence, the initial step involves identifying the exogenous factors to be included in the model. These factors should encompass financial, physical, geographical, legal, socio-cultural, and environmental aspects. Therefore, the following factors are considered: per capita income, cost of fixed assets, construction-related activity, gross national product (GNP) per capita, and unemployment rate.

To construct the model, further research must be conducted based on the following assumptions. Firstly, it is assumed that the level of investment activity in the region is determined by the amount of investment. Secondly, it is posited that the attractiveness of investment in the region is primarily influenced by the financial climate. The objective of this study is to develop an econometric model that can estimate the volume of fixed capital investment in the region, while considering the linear relationship between observed outcomes. It is assumed that the amount of fixed capital investment is dependent on several socio-economic indicators, which can be represented by a function. Thus, we aim to construct a function in the following form:

$$y_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}), \quad i = 1, \cdots, 31. \tag{2.1}$$

In the equation i – region; y_i – estimate of the volume of investments in the current year; x_{1i} – average per capita money income in the current year; x_{2i} – the cost of fixed assets in the current year; x_{3i} – the amount of work performed by the type of activity "Construction" in the current year; x_{4i} – GNP per capita in the current year; x_{5i} – unemployment rate in the current year (in percent) .

The Use of Logarithms in Econometric Model Construction

In the field of economics, it is common practice to employ logarithms when constructing econometric models, as this is believed to enhance the statistical properties of the estimated results. Failing to apply logarithms may lead to heteroskedastic errors, which can diminish the efficiency of least squares estimates. Consequently, drawing accurate statistical conclusions regarding the quality of the estimates becomes challenging. By taking logarithms, the variance is stabilized, resulting in constant error variance regardless of changes in the independent variable's value. Therefore, taking the natural logarithm of the dependent variable with respect to e becomes necessary to ensure the precision of the subsequent statistical analysis. The newly transformed function can be expressed as:

$$lny_i = f(x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}), \quad i = 1, \cdots, 31.$$

All computations for this study were conducted using the "RStudio" statistical data processing environment.

2.2.2 Regional Cluster Analysis of China: Unveiling Patterns and Relationships

Prior to constructing the assessment model, a hierarchical clustering analysis was conducted on the dataset, considering its limited number of observations. Ward's method was chosen as the clustering algorithm for this purpose. This method is based on the principle that an optimal classification should minimize the sum of squared differences within each group while maximizing the sum of squared differences between groups. In the context of this study, Ward's method proves to be particularly suitable for classifying the indicators.

The fundamental concept underlying Ward's method involves initially treating each of the n samples or variables as a separate cluster. Then, iteratively merging the clusters in such a way that the increase in the sum of squared deviations, denoted as S, is minimized with each merge. The two clusters that contribute the least to the increase in S are merged, and this process continues until all samples or variables are assigned to a single final cluster.

By employing Ward's method for hierarchical clustering analysis, we aim to identify meaningful patterns and relationships among the variables in our study. This procedure allows us to effectively group similar observations together and capture the inherent structure within the data. By applying this methodology, we can gain valuable insights into the underlying characteristics of the Chinese region under investigation.

In the process of dividing the dataset into k distinct classes, denoted as G_1, G_2, G_3 , and so on up to G_k , we measure the sum of squared deviations within each class. This can be represented by the equation:

$$S_t = \sum_{i=1}^{n_t} (X_{it} - \bar{X}_t)^T (X_{it} - \bar{X}_t),$$

Here, X_{it} represents the i-th sample in class G_t (an m-dimensional vector), n_t signifies the number of samples in G_t , and X_t denotes the centroid or center of gravity for class G_t . When merging two classes, say G_p and G_q, into a new combined class G_r, we consider three sums of squares: Sp, Sq , and Sr . The increase in the sum of squares, denoted as D^2_{pq}, is given by $D^2_{pq} = Sr - Sp - Sq$.

The magnitude of D_{pq}^2 provides insight into the reasonableness of merging the two classes. If G_p and G_q are closely related, D_{pq}^2 tends to be smaller, indicating a more justifiable classification. Conversely, if D_{pq}^2 is larger, it suggests an unreasonable classification. Thus, viewing the sum of squares of deviations resulting from merging two classes as a squared distance, we obtain the distance formula:

$$D_{pq}^2 = \frac{n_p n_q}{n_r} \left(X_p - \bar{X}_q\right)^T \left(X_p - \bar{X}_q\right).$$

Furthermore, we can utilize the recurrence formula to calculate the squared distance for merging class G_k with class G_r:

$$D_{kr}^2 = \frac{n_k + n_p}{n_r + n_k} D_{kp}^2 + \frac{n_k + n_q}{n_r + n_k} D_{kq}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2.$$

The Ward method adopts the Euclidean squared distance as its classification statistic. For any two samples i and j, the Euclidean squared distance is defined as:

$$\begin{aligned} d_{ij}^2 = &\left(X_{i1} - X_{j1}\right)^2 + \left(X_{i2} - X_{j2}\right)^2 + \cdots + \\ &+ \left(X_{i\ \mathrm{m}} - X_{j\ \mathrm{m}}\right)^2 = \sum_{n=1}^{\mathrm{m}} \left(X_{i\mathrm{n}} - X_{j\mathrm{n}}\right)^2. \end{aligned} \tag{2.2}$$

By employing these formulas and equations, the Ward method enables us to perform hierarchical clustering analysis based on the squared distances between samples, facilitating the identification of meaningful clusters within the dataset.

In equation 2.2, X_{in} and X_{jn} represent the n-th variable value for the i-th sample and the j-th sample, respectively. To mitigate the influence of variable magnitude on distance measurements between samples, a common practice in cluster analysis is to standardize the variables. By standardizing the variables, we transform them into standardized values, which are then utilized for conducting cluster analysis.

Standardization involves subtracting the mean value of each variable from its individual observations and dividing it by the standard deviation of that variable. This process ensures that all variables have a comparable scale, allowing for fair and meaningful comparison during clustering. By using standardized values, the impact of variables with larger magnitudes does not overshadow those with smaller magnitudes, as they are placed on a similar level playing field.

The standardization procedure enhances the reliability and interpretability of cluster analysis results. It enables accurate identification of patterns and relationships between samples based on their relative distances rather than absolute variable values. This approach ensures that the clustering algorithm focuses on the underlying structure and similarities among samples rather than being biased by differences in variable magnitudes.

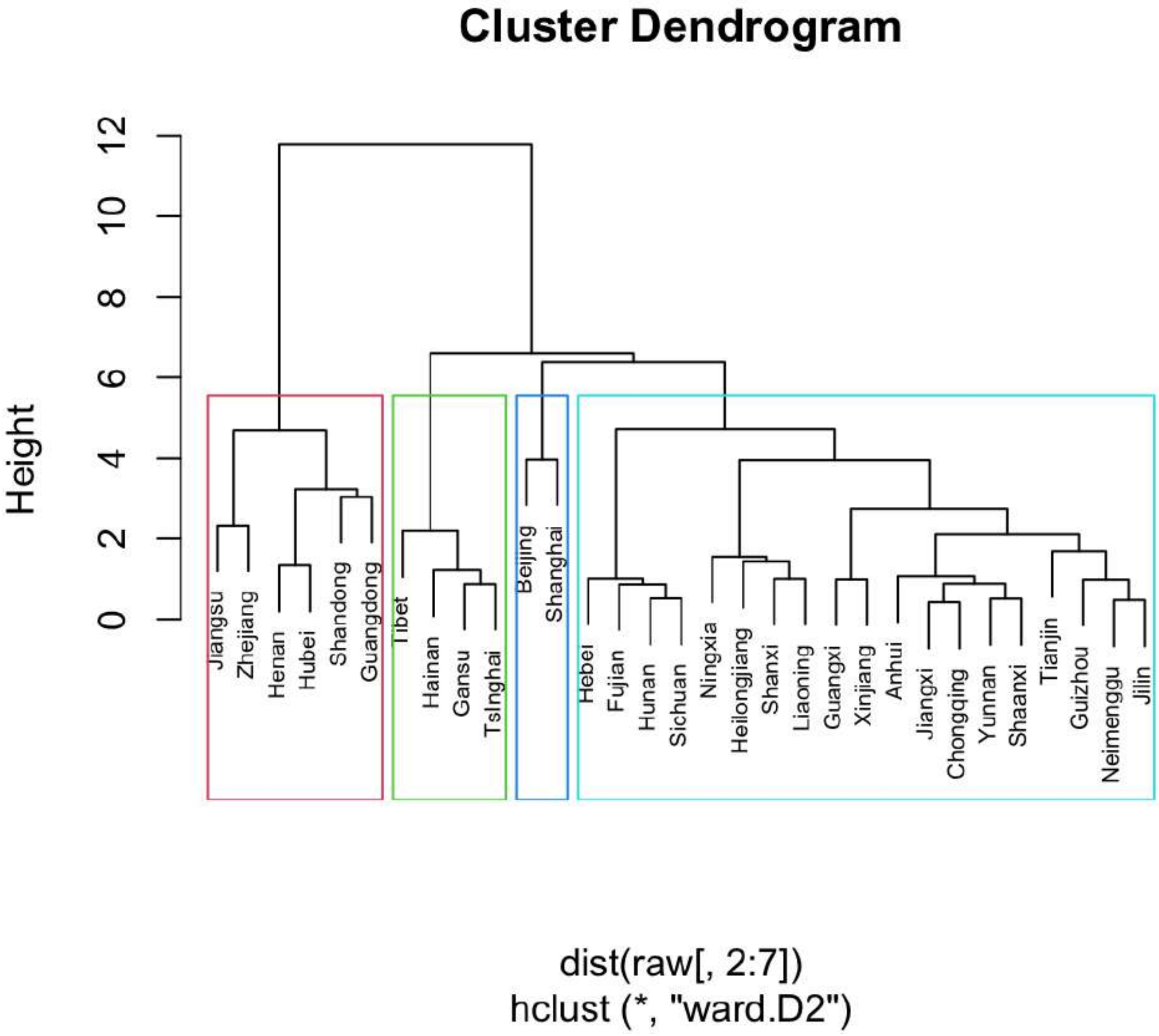


Figure 2.1: Cluster Dendrogram

The method is implemented in R as follows:

library(”factoextra”)

d < - dist(my data, method = “euclidean”)

res.hc < - hclust(d, method = “ward.D2”).

The findings of the calculations are visually depicted in Figure 2.1, presented as a tree diagram. The dataset utilized in this analysis incorporates data from reputable sources such as the World Bank and the China Statistical Yearbook . Inflation

Table 2.1: Descriptive statistics of clusters 1

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	2 693	20 854	10 087	35 478	2.5	10.36
1st Qu.	21 067	24 835	11 398	46 357	2.625	10.42
Median	28 380	33 730	12 434	62 201	2.750	10.62
Mean	26 116	31 348	16 920	63 335	2.833	10.63
3rd Qu.	34 519	37 118	23 775	82 561	2.950	10.84
Max.	42 046	39 658	27 957	89 705	3.40	10.92

Table 2.2: Descriptive statistics of clusters 2

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	11 547	1 376	147.9	1 311	2.3	7.589
1st Qu.	12 395	2 456	279.1	2 296	2.6	8.096
Median	17 506	3 310	364.8	3 544	2.7	8.309
Mean	14 778	2 973	675.8	3 965	2.7	8.219
3rd Qu.	19 889	3 828	761.6	5 212	:2.8	8.433
Max.	22 553	3 897	1 825.4	7 460	3.1	8.670

Table 2.3: Descriptive statistics of clusters 3

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	57 230	10 946	6 426	28 015	1.400	8.888
1st Qu.	57 669	11 258	7 254	28 669	2.025	8.924
Median	58 109	11 570	8 082	29 324	2.650	8.960
Mean	58 109	11 570	8 082	29 324	2.650	8.960
3rd Qu.	58 548	11 881	8 909	29 978	3.275	8.996
Max.	58 988	12193	9 737	30 633	3.900	9.032

Table 2.4: Descriptive statistics of clusters 4

parameter	x_1	x_2	x_3	x_4	x_5	y
Min.	16 704	3 807	549.2	3 444	2.200	8.224
1st Qu.	20 571	10 003	2 675.8	15 999	3.250	9.366
Median	21 484	10 467	4 726.4	19 425	3.500	9.772
Mean	24 843	12 285	5 074.6	21 459	3.479	10.130
3rd Qu.	25 183	14 951	6 628.0	28 825	3.900	10.130
Max.	58 988	19 083	11 400.3	36 980	4.200	10.417

has been accounted for by adjusting all values to comparable prices. The figure illustrates the division of the data into distinct clusters, each represented by a unique color. In this particular analysis, the data has been partitioned into four clusters, providing a hierarchical representation of its underlying structure. This hierarchical characteristic is visually evident upon examination of the results. Furthermore, the descriptive statistics associated with the clusters lend further support to this conclusion.

The subsequent analysis will delve into each cluster individually, providing insights into their respective characteristics.

Group 1 comprises six regions characterized by substantial investment volumes. Although there are discernible differences in the values of individual factors among these regions, the disparities in investment attractiveness differentials between them do not reach statistical significance. Descriptive statistics pertaining to the classification of the indicators under study are presented in Table 2.1.

Group 2 consists of four regions exhibiting comparatively lower levels of investment. The descriptive statistics within this group demonstrate a higher level of consistency when compared to Group 1.

Group 3 comprises two regions, namely Beijing and Shanghai, which stand out as the most prosperous areas nationwide. The disparities between these two regions in each factor are relatively minor.

Group 4 is the largest cluster, encompassing nineteen regions. Not only do the regions in this group exhibit significant variations in terms of investment amounts, but their other factors also differ significantly.

By conducting a detailed analysis of each cluster, we gain valuable insights into the distinct characteristics and patterns observed within the data. Such an approach allows for a comprehensive understanding of the regional dynamics and investment landscapes across the studied areas.

The determination of the appropriate number of clusters, denoted as k, holds significant importance when applying the "k-means" method. To assess the relevance of selecting k, we employ the "elbow method" by plotting the dependence of intra-group scattering against the number of clusters. As depicted in Figure 2.2, a noticeable decrease in intra-group scattering occurs at $k = 2$, followed by stabilization at $k = 4$. This analysis indicates that dividing the clusters into four groups aligns more consistently with the geographic and economic characteristics of the Chinese regions.

Based on these findings, the Chinese regions were categorized into four distinct clusters, differing in terms of investment levels and geographical attributes. The cluster analysis results reveal a discernible relationship between the investment attractiveness of regions and various factors, including per capita income, fixed asset costs, gross national product (GNP), construction activity, and unemployment rate.

Consequently, employing multiple regression analysis techniques to develop models based on observations within these clusters holds meaningful implications.

By utilizing multiple regression analysis, we can establish predictive models that consider the aforementioned factors to understand their influence on the investment attractiveness of different regions. These models hold potential for uncovering valuable insights and enabling informed decision-making in the context of regional development and investment strategies.

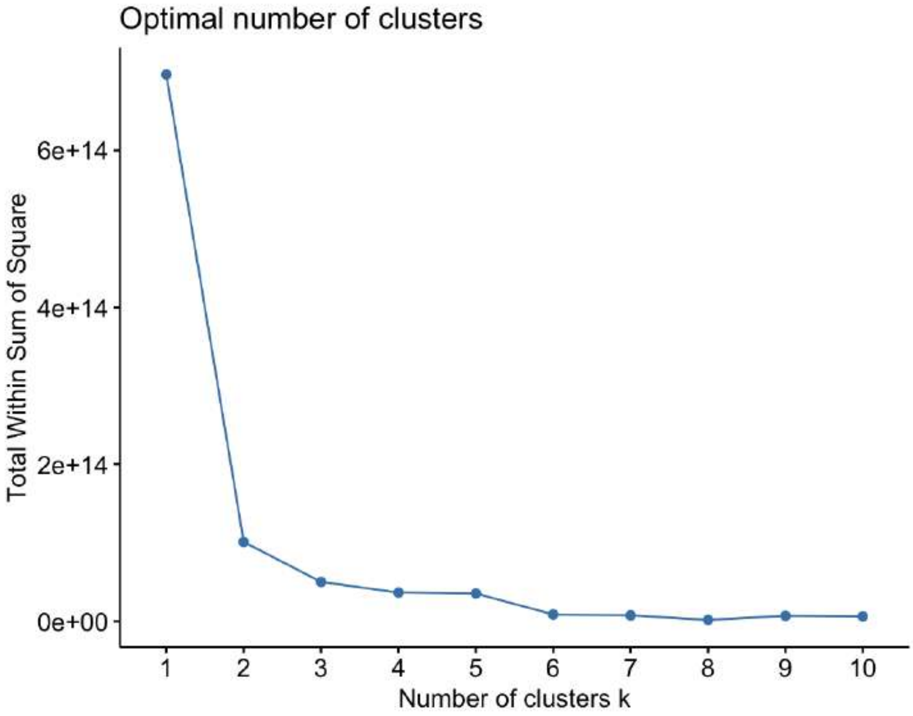


Figure 2.2: Choice relevance k –number of clusters

2.2.3 Building econometric models

In order to comprehensively analyze large clusters, a multiple regression analysis will be conducted for each year of available data. This approach allows us to generate models that capture the relationships between various factors and provide a holistic evaluation. By examining these models collectively, we can draw general conclusions regarding the influence of the considered factors on investment attractiveness.

Considering that separate models will be developed for each group of observations within the clusters, it is reasonable to assume a linear relationship between the factors. The linear regression framework provides a systematic and quantitative means to assess the impact of independent variables on the dependent variable of interest. This assumption aids in interpreting the coefficients of the regression models, enabling the identification of the magnitude and direction of the influence exerted by each factor.

For smaller clusters, where the number of observations may be limited, correlation analysis offers valuable insights into the degree of influence exhibited by the factors. Correlation analysis helps ascertain the strength and direction of the linear relationship between pairs of variables. By examining the correlation coefficients, we can gauge the extent to which changes in one variable correspond to changes in another, thereby evaluating the degree of association between the factors.

By employing multiple regression analysis for large clusters and correlation analysis for small clusters, we adopt appropriate statistical techniques tailored to the characteristics of each cluster size. These analytical approaches facilitate a comprehensive examination of the interplay between factors and their impact on investment attractiveness within different contexts.

Multiple regression model for group 4

The analysis focused on Group 4, consisting of 19 regions, which were selected as the primary sample for the study. The data collected from these 19 regions was used to establish an observational model, enabling us to examine the relationship between various factors and investment attractiveness within this group.

The observational model employed in this study captured the interplay between independent variables and investment attractiveness as the dependent variable. By leveraging the dataset from each of the 19 regions within Group 4, we aimed to identify significant factors that contribute to variations in investment attractiveness across these regions.

Through this observational model, we sought to uncover insights into the complex dynamics and determinants of investment attractiveness within the selected regions. This approach allowed us to explore the unique characteristics and factors driving investment patterns and develop a more nuanced understanding of their influence.

By utilizing the data from all 19 regions in Group 4, we aimed to provide a comprehensive analysis that takes into account the diverse economic, geographical, and social contexts of these regions. This approach enables us to draw meaningful conclusions and recommendations regarding investment strategies and regional development initiatives based on the observed relationships between factors and investment attractiveness within this specific group.

$$\ln y_{2017i} = \alpha + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i} + \beta_4 \cdot x_{4i} + \beta_5 \cdot x_{5i} + \epsilon_i, \tag{2.3}$$

where i – region; ϵ_i – the total effect of factors not taken into account by the model.

The estimates of the regression equation 2.3 were found and the results are presented in Table 2.5 .

Table 2.5: Coefficient estimates of the group 4

coefficient	Estimate	Std. Error	t-value	$Pr(>\|t\|)$	Signif.
$\hat{\alpha}$	1.005e+01	4.828e-01	20.817	2.29e-11	*** [1]
$\hat{\beta}_1$	-2.508e-05	8.018e-06	-3.128	0.00801	** [2]
$\hat{\beta}_2$	6.247e-05	3.399e-05	1.838	0.08901	. [3]
$\hat{\beta}_3$	3.220e-05	3.968e-05	0.812	0.43167	
$\hat{\beta}_4$	2.418e-05	1.956e-05	1.236	0.23822	
$\hat{\beta}_5$	-3.442e-01	1.334e-01	-2.580	0.02284	* [4]

[1] A p-value less than 0.001 indicates very strong evidence against the null hypothesis.

[2] A p-value less than 0.01 indicates even stronger evidence against the null hypothesis.

[3] A p-value less than 0.1 is considered weak evidence against the null hypothesis.

[4] A p-value less than 0.05 indicates strong evidence against the null hypothesis and we can reject it in favor of the alternative hypothesis.

Therefore, the regression equation is as follows:

$$\begin{aligned} \ln y_{\hat{2017i}} =& 10.05 - 0.00002508 \cdot x_{1i} + 0.00006247 \cdot x_{2i} + 0.0000322 \cdot x_{3i} + \\ &+ 0.000002.418 \cdot x_{4i} - 0.3442 \cdot x_{5i}, \end{aligned}$$

According to the t–test, three of the five coefficient estimates ($\hat{\alpha}$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_5$) are statistically significant at the 5 % level of significance, with their p–value being less than 0.1. The values of R^2 and R^2_{adj} are 0.86 and 0.8062, respectively, indicating that approximately 80 % of the variation in the dependent variable is explained by the

R^2	0.86
R^2_{adj}	0.8062
F	15.98
p–value(F)	3.756e-05

regression. Furthermore, the overall quality of the model is sufficient, as evidenced by Fisher's criteria: $F - statistics = 15.98$, with a corresponding $p - value = 3.756e - 05$, which is less than 0.05 and close to zero. This result confirms that the average quality of the whole model is satisfactory.

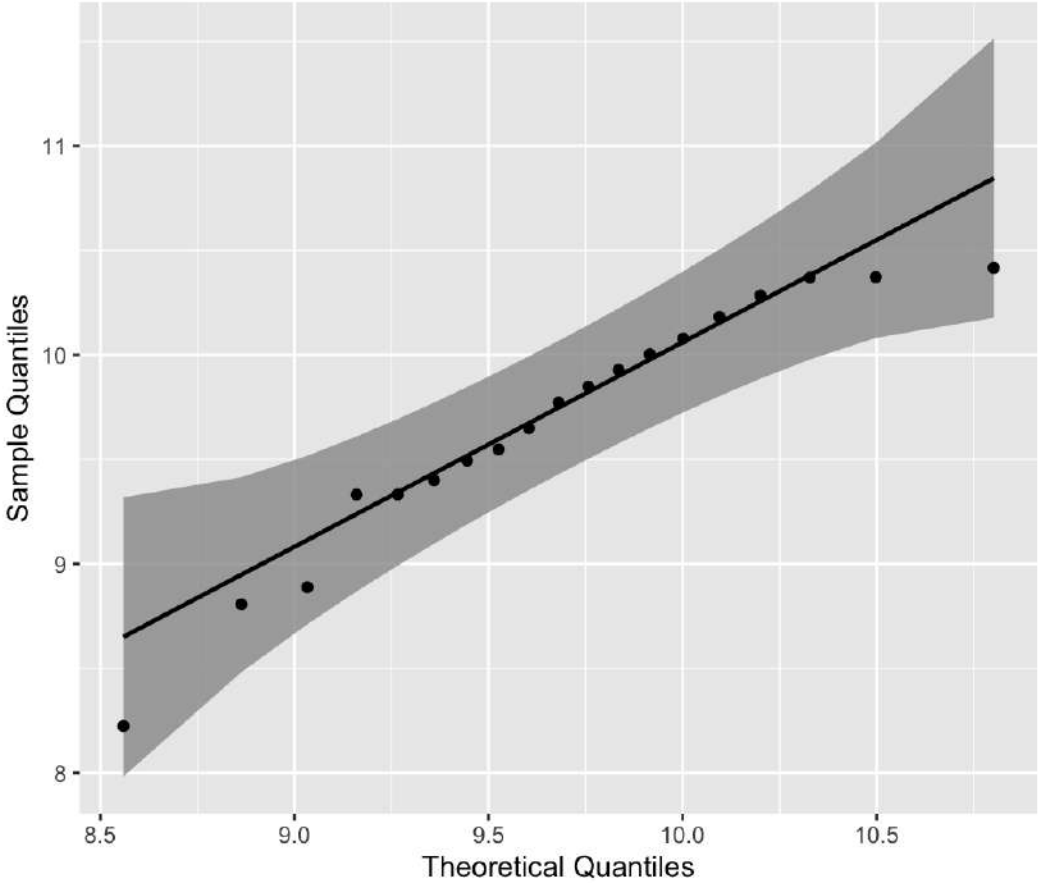


Figure 2.3: plot for the group 4 residuals.

In order to assess the normality of the residual distribution, it is crucial to examine the QQ plot (Figure 2.3). The QQ plot provides a graphical representation that allows us to compare the observed quantiles of the residuals against the expected quantiles under a theoretical normal distribution.

Upon analyzing the QQ plot, it can be observed that the majority of points closely align along a single line, suggesting that the residual distribution follows a normal distribution. However, it is worth noting the presence of some outliers, which deviate from the expected pattern. These outliers could indicate potential departures from normality in certain regions of the residual distribution.

The normality of the residual distribution is an important assumption for many statistical models and tests. A departure from normality may indicate the presence of influential observations or violations of model assumptions. Therefore, it is necessary to carefully evaluate and interpret these outliers in relation to the overall characteristics of the data.

While the majority of the points on the QQ plot conform to the expected normal distribution pattern, the existence of outliers raises a degree of caution. Further diagnostics and assessments should be conducted to investigate the nature and impact of these outliers on the model's validity. This evaluation of the residual distribution's normality aids in ensuring the reliability and robustness of the statistical analysis conducted in this study.

lag	Autocorrelation	DW Statistic	p–value
1	0.15803	1.665274	0.194

χ^2	p-value(Breusch-Pagan)
2.945036	0.086142

To validate the accuracy of the selected factors included in the model, an assessment of the autocorrelation of the residuals was performed. The Durbin-Watson criterion was utilized for this purpose, resulting in a value of $DW = 1.6653$. This value suggests that the residuals exhibit no significant autocorrelation, reinforcing the appropriateness of the factor selection within the model.

However, it was imperative to further examine the homoscedasticity of the residuals. To achieve this, the Breusch-Pagan test was employed, yielding a test statistic of $\chi^2 = 2.945036$ and a corresponding $p - value = 0.0861$. While the p-value does not reach conventional levels of statistical significance, it provides a moderate indication that the residuals may display homoscedasticity. Hence, based on this test, there is limited evidence to suggest heteroscedasticity within the residuals.

The analysis of the regression residuals overall yielded satisfactory results, as both the autocorrelation and heteroscedasticity tests did not identify any major violations of assumptions. However, it is important to acknowledge that these tests have their limitations and should be interpreted in conjunction with other diagnostic measures to ensure the reliability of the model.

The Table 2.6 shows the Variance Inflation Factor (VIF) for each variable in the model. VIF is a measure of multicollinearity, which quantifies how much the variance of a regression coefficient is inflated due to the correlation with other predictors.

Table 2.6: Multicollinearity Assessment - VIF Values

Features	VIF Factor
const	0.003504
x_1	0.049529
x_2	0.015893
x_3	0.009322
x_4	0.008444
x_5	0.027871

The "Features" column represents the names of the variables included in the model. The "VIF Factor" column displays the calculated VIF values for each variable.

In the table, we observe the following:

The constant term (intercept) has a VIF Factor of 0.003504, indicating that it has very low collinearity with other predictors. Variable x_1 has a VIF Factor of 0.049529, which suggests only moderate collinearity. Variables x_2, x_3, x_4, and x_5 have VIF Factors of 0.015893, 0.009322, 0.008444, and 0.027871 respectively, indicating relatively low levels of collinearity. Overall, the VIF factors are all relatively low, suggesting that there is no severe multicollinearity issue among the variables in the model. This indicates that the variables can be considered as independent and their coefficients can be reliably interpreted in the context of the academic study.

These findings provide evidence of acceptable collinearity levels, supporting the suitability of the variables for inclusion in this paper.

To obtain the final model, a stepwise regression algorithm was employed. This algorithm systematically identifies the factor(s) with the least significant coefficient (maximum p-value) in each step and removes them from the model. By iteratively applying this procedure, the algorithm progressively refines the model until reaching a point where all factors possess statistically significant coefficients. Employing the stepwise regression approach helps streamline the model by excluding less influential factors and improving its interpretability and predictive performance.

$$\ln y_{\hat{2017}i} = 9.666 - 0.00001741 \cdot x_{1i} + 0.000126 \cdot x_{2i} - 0.3164 \cdot x_{5i};$$

Apply the same process to the data for other years 2009 – 2016:

$$2009 : \ln \hat{y_{2009i}} = 7.473 + 0.0001078 \cdot x_{2i} + 0.00000001253 \cdot x_{3i} + 0.00003056 \cdot x_{4i};$$

$$2010 : \ln \hat{y_{2010i}} = 7.988 - 0.00003046 \cdot x_{1i} + 0.0001484 \cdot x_{2i} + 0.00000001938 \cdot x_{3i};$$

$$2011 : \ln \hat{y_{2011i}} = 8.539 + 0.000000003164 \cdot x_{3i} + 0.00006634 \cdot x_{4i} - 0.1231 \cdot x_{5i};$$

$$2012 : \ln \hat{y_{2012i}} = 9.025 - 0.00001169 \cdot x_{1i} + 0.00006653 \cdot x_{4i} - 0.1351 \cdot x_{5i};$$

$$\begin{aligned} 2013 : \ln \hat{y_{2013i}} = & 9.452 - 0.00001755 \cdot x_{1i} + 0.00001159 \cdot x_{2i} + 0.00005484 \cdot x_{4i} - \\ & - 0.1679 \cdot x_{5i}; \end{aligned}$$

$$2014 : \ln \hat{y_{2014i}} = 9.677 - 0.00001669 \cdot x_{1i} + 0.00005619 \cdot x_{4i} - 0.229 \cdot x_{5i};$$

$$\begin{aligned} 2015 : \ln \hat{y_{2015i}} = & 9.807 - 0.00002218 \cdot x_{1i} + 0.000000003823 \cdot x_{3i} + 0.00003945 \cdot x_{4i} - \\ & - 0.16768 \cdot x_{5i}; \end{aligned}$$

$$2016 : \ln \hat{y_{2016i}} = 9.122 - 0.00002709 \cdot x_{1i} + 0.00007275 \cdot x_{2i} + 0.00000000645 \cdot x_{3i}.$$

Multiple regression model for group 1

A multiple regression model was constructed for the group 1 using the analysis method in Table 2.7.

Table 2.7: Coefficient estimates of group 1

coefficient	Estimate	Std. Error	t–value	Pr(>|t|)	Signif.
$\hat{\alpha}$	9.130e+00	4.838e-01	18.873	0.0337	*
$\hat{\beta}_1$	-5.831e-06	4.699e-06	-1.241	0.4318	
$\hat{\beta}_2$	4.330e-05	1.367e-05	3.168	0.1947	
$\hat{\beta}_3$	1.585e-05	5.204e-06	3.045	0.2020	
$\hat{\beta}_4$	-9.283e-06	4.977e-06	-1.865	0.3133	
$\hat{\beta}_5$	2.178e-01	1.000e-01	2.178	0.2741	

Therefore, the regression equation is as follows:

$$\begin{aligned}\ln y_{2\hat{0}17i} =& 9.13 - 0.000005831 \cdot x_{1i} + 0.0000433 \cdot x_{2i} + 0.00001585 \cdot x_{3i} - \\ & - 0.000009283 \cdot x_{4i} + 0.2178 \cdot x_{5i},\end{aligned}$$

The statistical significance of the coefficients obtained from the model is presented in Table 6. We observe that only the estimates of the free coefficients demonstrate statistical significance at the 5% level.

To comprehensively evaluate the overall quality of the model, we turn our attention to the first cohort. This assessment allows us to gauge the extent to which the model captures and explains the observed data for this specific group.

Based on our analysis, it can be affirmed that the average quality of the entire model is deemed satisfactory. This conclusion is drawn upon considering various factors such as the statistical significance of the coefficients, the fit of the model to the observed data, and other relevant performance metrics.

By establishing a model that encompasses significant and influential factors, we are able to gain valuable insights into the relationship between these factors and the outcome of interest. This understanding contributes to a more comprehensive understanding of the underlying dynamics and mechanisms at play within the first cohort. Additionally, it serves as a foundation for further exploration and refinement of the model, ultimately leading to enhanced predictive capabilities and decision-making.

The F-statistic corresponding to the model is statistically significant with a p-value less than 0.05, indicating that the model is acceptable at a 5 % level of signif-

R^2	0.9824
R^2_{adj}	0.8946
F	11.19
$p\text{–}value(F)$	0.0223

icance. The model demonstrates a high degree of explanatory power, as it explains approximately 90 % of the variation in the outcome variables.

lag	Autocorrelation	D-W Statistic	p-value
1	-0.4205415	2.354857	0.162

χ^2	p-value(Breusch-Pagan)
0.1263405	0.07222

The Durbin-Watson statistic (D-W) of 2.35 falls within the range of $1.5 < DW < 2.5$, indicating that the residuals in the model for the first group do not exhibit significant autocorrelation. This result suggests that the independence assumption of the regression analysis is met, further supporting the validity of the findings.

Additionally, the Breusch-Pagan test, which assesses the homoscedasticity of the residuals, yields a p-value greater than 0.05. This indicates that there is no strong evidence to reject the null hypothesis of residuals being homoscedastic. While the test does not provide definitive proof of homoscedasticity, it suggests that any deviations from homoscedasticity are likely minor and do not significantly impact the overall quality of the model.

Based on these statistical analyses, we can conclude that the model quality for the first group is considered normal. The results indicate that the assumptions underlying the model, such as the absence of residual autocorrelation and relatively minor deviations from homoscedasticity, are reasonably satisfied within this context.

The Table 2.8 provides the Variance Inflation Factor (VIF) values for the variables included in the model.

The constant term (denoted as "const") has a VIF value of 0.01198953, indicating that it exhibits minimal collinearity with the other predictors. This suggests that the constant term does not have a substantial impact from correlations with the remaining variables.

Variable "x_1" displays a VIF value of 0.088700, indicating some level of collinearity with other predictors. Although there is a moderate degree of correlation, it remains within an acceptable range.

Table 2.8: Multicollinearity Assessment - VIF Values

Features	VIF Factor
const	0.011989
x_1	0.088700
x_2	0.060412
x_3	0.032016
x_4	0.025576
x_5	0.020602

Variables "x_2," "x_3," "x_4," and "x_5" exhibit VIF values of 0.0604120, 0.03201671, 0.0255766, and 0.0206025, respectively. These values suggest relatively low levels of collinearity with the other predictors.

Collectively, the VIF values indicate that there is no severe multicollinearity issue among the variables considered in the model. This implies that the variables are reasonably independent, allowing for reliable interpretation of their coefficients within this study.

To derive the final model, a stepwise regression algorithm was employed. This iterative procedure systematically identifies the factor with the least significant coefficient (maximum p-value) at each step and removes it from the model. By iteratively eliminating less influential factors, the algorithm refines the model and selects the most relevant variables to be included. This approach helps streamline the model by focusing on the statistically significant factors, enhancing its interpretability and predictive performance.

The utilization of the stepwise regression algorithm allows us to identify the key factors that contribute significantly to the observed outcome in the first group. By including only the most relevant variables, the final model effectively captures the essential information and relationships within this specific cohort.

$$\ln y_{\hat{2017}i} = 9.179 + 0.0000371 \cdot x_{2i} - 0.000005542 \cdot x_{4i} + 0.2202 \cdot x_{5i},$$

Apply the same process to the data for other years.

$$2009 : \ln y_{\hat{2009}i} = 8.518 + 0.00009.835 \cdot x_{2i} - 0.00001247 \cdot x_{4i};$$

$$2010 : \ln y_{\hat{2010i}} = 6.261 - 0.0001188 \cdot x_{1i} - 0.00006069 \cdot x_{2i} + 0.000000007946 \cdot x_{3i};$$

$$2011 : \ln y_{\hat{2011i}} = 8.072 + 0.00006173 \cdot x_{2i} + 0.1711 \cdot x_{5i};$$

$$2012 : \ln y_{\hat{2012i}} = 8.195 + 0.00004987 \cdot x_{2i} + 0.000000001065 \cdot x_{3i} + 0.1992 \cdot x_{5i};$$

$$2013 : \ln y_{\hat{2013i}} = 8.272 + 0.00004462 \cdot x_{2i} + 0.000000001145 \cdot x_{3i} + 0.231 \cdot x_{5i};$$

$$2014 : \ln y_{\hat{2014i}} = 9.788 - 0.00002668 \cdot x_{1i} + 0.00002101 \cdot x_{2i} + 0.000000002385 \cdot x_{3i};$$

$$2015 : \ln y_{\hat{2015i}} = 10 - 0.00002476 \cdot x_{1i} + 0.00002912 \cdot x_{2i} + 0.000000002255 \cdot x_{3i};$$

$$2016 : \ln y_{\hat{2016i}} = 10.19 - 0.00002719 \cdot x_{1i} + 0.00002742 \cdot x_{2i} + 0.000000002265 \cdot x_{3i}.$$

2.2.4 Cluster-specific examinations

To assess the influence of various factors on investment attractiveness, multiple regression analysis was selected as the primary method in this study. This analytical approach allows for a comprehensive examination of the relationships between independent variables and the dependent variable of interest. However, considering the presence of a hierarchical structure within the dataset, it was essential to address this issue to ensure accurate analysis.

To mitigate the potential effects of the hierarchical structure, the observations were divided into four distinct groups. This division facilitated a more refined analysis by analyzing each group separately. By doing so, we aimed to account for the unique characteristics and dynamics present within each group, thereby reducing any potential bias or confounding factors that may arise from the hierarchical structure.

It is worth noting that for small clusters, multiple regression analysis may not be feasible due to limited sample size and insufficient statistical power. As an alternative, correlation analysis was employed to assess the degree of influence exhibited by the factors within these smaller clusters. Correlation analysis provides insights into the strength and direction of the linear relationship between pairs of variables, enabling us to evaluate the magnitude of their association.

By employing multiple regression analysis for large clusters and correlation analysis for small clusters, we adopted appropriate statistical techniques tailored to the characteristics of each cluster size. This methodology allowed us to effectively explore and measure the influence of factors on investment attractiveness, while accounting for any hierarchical structure and limitations inherent in small-sample analyses. Overall, this approach enhances the rigor and validity of the study's findings and contributes to a more comprehensive understanding of the factors impacting investment attractiveness.

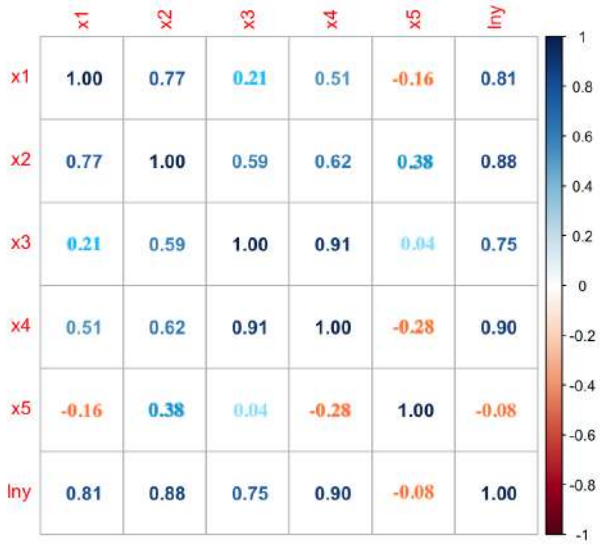
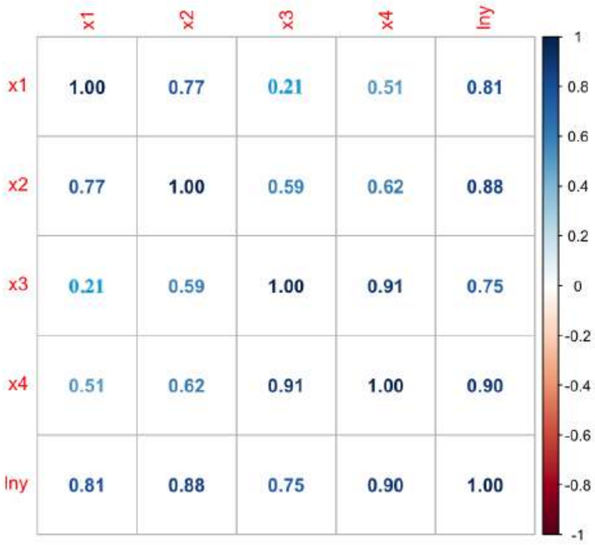


Figure 2.4: Thermal correlation map for 2-group

To conduct a statistical examination of the second and third groups, correlation analysis was utilized due to the small number of clusters involved.

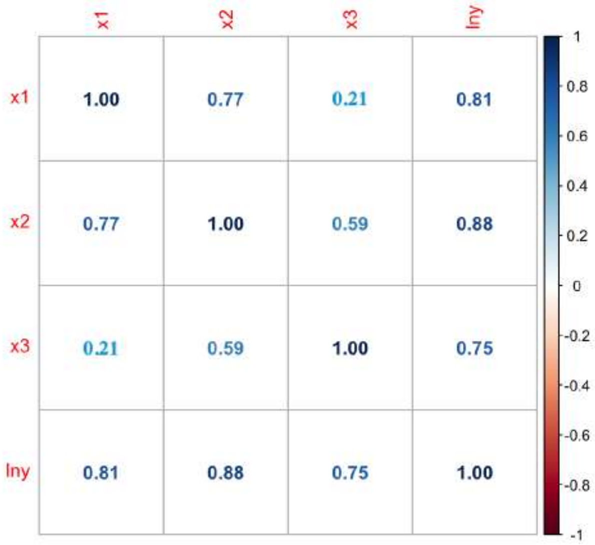
For the second group, a corrplot correlation heat map was generated (Figure 2.4) to visually represent the correlations between variables. Notably, the correlation between variable x_5 and the outcome variable is weak ($r = 0.08$). Consequently, it



	x1	x2	x3	x4	lny
x1	1.00	0.77	0.21	0.51	0.81
x2	0.77	1.00	0.59	0.62	0.88
x3	0.21	0.59	1.00	0.91	0.75
x4	0.51	0.62	0.91	1.00	0.90
lny	0.81	0.88	0.75	0.90	1.00

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

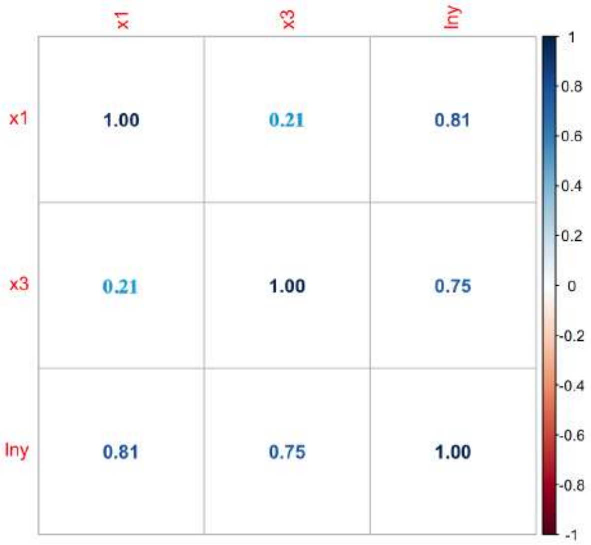
Figure 2.5: Map after remove x5



	x1	x2	x3	lny
x1	1.00	0.77	0.21	0.81
x2	0.77	1.00	0.59	0.88
x3	0.21	0.59	1.00	0.75
lny	0.81	0.88	0.75	1.00

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 2.6: Map after remove x5, x4



	x1	x3	lny
x1	1.00	0.21	0.81
x3	0.21	1.00	0.75
lny	0.81	0.75	1.00

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 2.7: Map after remove x5, x4, x2

is logical to exclude x_5 from the factor set. After removing x_5 , the correlation heat map takes on a modified form (Figure 2.5).

As depicted in Figure 5, x_4 exhibits a strong correlation with $x_3 (r > 0.7)$, as well as moderate correlations with x_1 and x_2 $(0.4 < r < 0.7)$. Therefore, it is reasonable to exclude x_4 from the factor set, despite its stronger correlation with the outcome variable (Figure 2.6).

In the subsequent step, the decision is made to remove x_2 due to its moderate correlation with both x_1 and x_3, while the correlation between x_1 and x_3 remains notably low. The resulting outcomes are then obtained (Figure 2.7).

This sequential elimination of covariates within the explanatory variables allows for the identification of the most strongly correlated factors. In this case, only x_1 and x_3 remain as independent variables. However, it should be noted that in the second cluster, x_3 exhibits a weaker association with the outcome variable y compared to x_1 . Thus, it can be concluded that within the second cluster, the primary factor influencing the formation of y values is the factor x_1.

Regarding the third cluster, due to the limited amount of data, it is not possible to present a detailed correlation heat map. Consequently, there is insufficient information available to draw any meaningful conclusions about the statistical relationship between the studied indicators and investment attractiveness within this cluster. The small number of observations in this cluster likely indicates a deviation from the overall pattern, highlighting the need for further investigation and caution when interpreting results within this specific group.

2.3 Discussion and Analysis

The chosen approach of employing multiple regression analysis to evaluate the influence of various factors on investment attractiveness allowed for a comprehensive examination of the relationships between independent variables and the dependent variable. To account for the heterogeneity in cluster sizes, the observations were divided into four distinct groups, and separate analyses were conducted for each cluster.

In the first group, representing the most attractive areas for investment, the results of the multiple regression analysis revealed a strong correlation between the volume of investment and the cost of fixed assets (x_2), as well as the amount of work performed by the type of activity "Construction" (x_3). Notably, the correlation with the cost of fixed assets (x_2) was identified as the strongest. This finding implies that in these highly attractive regions, investors tend to allocate a significant portion of their resources towards acquiring costly fixed assets, which positively influences investment volume.

Conversely, in the fourth group characterized by low investment attractiveness, a strong correlation was observed between investment volume and average per capita money income (x_1), ,as well as the amount of work performed by the type of activity "Construction"(x_3). However, the correlation with average per capita money income(x_1) was identified as the primary factor influencing investment volume within this group. This suggests that in less attractive regions, lower levels of average per capita income may hinder investment opportunities, leading to reduced investment volumes.

For the second group, which exhibited the lowest investment attractiveness, the correlation analysis conducted due to the small cluster size indicated that investment volume was closely correlated with average per capita money income (x_1)and the amount of work performed by the type of activity "Construction"(x_3). Here too, average per capita money income (x_1) emerged as the primary factor influencing investment volume. These findings suggest that in regions with low investment attractiveness, the level of average per capita income plays a critical role in attracting investment.

In light of these results, it is evident that analyzing the investment attractiveness of regions separately and considering their respective economic characteristics is crucial. This approach allows policymakers to develop targeted strategies and policies tailored to the specific needs and challenges faced by each region. By understanding the factors that influence investment attractiveness on a regional level, policymakers can foster conditions conducive to increased investment and economic growth.

It is important to acknowledge the limitations of this study, such as the potential presence of other unmeasured factors that could contribute to investment attractiveness. Additionally, the generalizability of the findings should be approached with caution due to the specific context and characteristics of the regions under investigation. Further research is needed to validate and expand upon these findings, particularly through the inclusion of additional variables and the consideration of broader economic and social factors influencing investment decisions.

2.4 Conclusion to Chapter 2

In conclusion, this study employed multiple regression analysis to examine the influence of various factors on investment attractiveness. By clustering the observations into four distinct groups and conducting separate analyses for each cluster, a comprehensive understanding of the relationships between independent variables and investment volume was achieved.

The findings revealed important insights into the factors influencing investment attractiveness within different clusters. In highly attractive regions (first group), the cost of fixed assets emerged as a strong determinant of investment volume, indicating that investors in these areas allocate significant resources towards acquiring costly assets. Conversely, in regions characterized by low investment attractiveness (fourth group), average per capita money income was identified as the primary factor influencing investment volume. This suggests that lower levels of average per capita income in these regions may hinder investment opportunities.

For the second group, which displayed the lowest investment attractiveness, both average per capita money income and the amount of work performed in the "Construction" sector were closely correlated with investment volume. However, average per capita money income was found to be the primary factor shaping investment patterns. These results highlight the significance of considering regional economic characteristics when assessing investment attractiveness and developing targeted policies.

It is important to acknowledge the limitations of this study, including the potential influence of unmeasured factors on investment attractiveness. Furthermore, caution should be exercised when generalizing the findings to other contexts due to the specific characteristics of the regions under investigation. Future research should expand upon these findings by incorporating additional variables and examining broader economic and social factors that might impact investment decisions.

By comprehensively analyzing investment attractiveness on a regional level and considering specific economic characteristics, policymakers can develop tailored strategies to enhance investment opportunities and foster economic growth. Understanding the factors driving investment decisions can inform effective policy development and create an environment conducive to increased investment in different regions. Further research is necessary to validate these findings and provide a more comprehensive understanding of investment attractiveness in diverse contexts.

Based on these results, several recommendations and suggestions can be made. Firstly, policymakers should prioritize investments in regions with high attractiveness by focusing on the acquisition of costly fixed assets. This can be achieved through the provision of incentives and support for businesses looking to invest in these regions.

Secondly, in regions characterized by low investment attractiveness, efforts should be directed towards improving average per capita income levels. This can be accomplished by implementing targeted initiatives aimed at creating employment opportunities, enhancing skills training programs, and promoting entrepreneurship and innovation. Increasing average per capita income will likely lead to a more conducive investment environment.

Furthermore, policymakers should closely monitor the construction industry's performance as it emerged as a significant factor influencing investment volume across different clusters. Identifying potential obstacles or inefficiencies within the construction sector and addressing them can contribute to increased investment attractiveness.

For future development, it is imperative to conduct further research to validate and expand upon these findings. This could involve exploring additional variables such as infrastructure development, access to financing, and government policies. Broadening the scope of analysis to include social and cultural factors could also provide valuable insights into investment decisions.

Moreover, ongoing monitoring and evaluation of investment attractiveness are essential to adapt policies based on changing economic conditions and market dynamics. Regular assessments can help identify emerging trends and challenges, allowing policymakers to make informed decisions and adjust strategies accordingly.

Chapter 3

Analyzing and Simulating Air Quality Index using Stepwise Regression: Exploring Trends and Evaluating Fit

This chapter extensively utilizes the stepwise regression method as the primary analytical technique to investigate and simulate the factors affecting the Air Quality Index (AQI). Subsequently, a meticulous evaluation is conducted by comprehensively comparing the models derived from the stepwise regression process based on the Akaike Information Criterion (AIC), with the aim of identifying the model characterized by the smallest AIC. Furthermore, a comparative analysis is performed between the final stepwise regression model and the model selected using the AIC criterion. This rigorous methodology enables a systematic examination of the diverse factors influencing AQI, contributing to a deeper understanding of the intricacies involved and offering valuable insights into the development of effective strategies for air quality management. Some algorithms discussed in this chapter have been utilized by the author in their published works [100] and [101].

3.1 Data

3.1.1 Data Source and Collection

This study relies on data from the National Bureau of Statistics of China as the source of all independent variables, ensuring the credibility and authority of the data. The National Bureau of Statistics serves as a vital governmental agency responsible for collecting, organizing, and disseminating a broad range of macroeconomic and social statistics. These statistics originate from diverse sources, including government surveys, sample surveys, and censuses, thus establishing high levels of reliability and wide coverage.

The dependent variable data, on the other hand, is derived through the utilization of the national air quality calculation platform. This platform operates as a specialized tool jointly managed by environmental protection departments and relevant scientific research institutions to monitor and assess air quality across China. Employing advanced sensing technology and monitoring equipment, the platform acquires real-time data on various air quality indicators such as PM2.5, PM10, sulfur dioxide, nitrogen dioxide, among others. These indicators serve as vital parameters for evaluating air quality and are capable of reflecting pollution levels within different provinces during specific time periods.

To ensure comprehensive datasets, air quality data spanning from 2013 to 2017 were collected, encompassing all provinces throughout China. The selected seven-year timeframe provides a relatively long-term perspective, facilitating the analysis of air quality trends and disparities between distinct years and regions. Furthermore, considering variations in development levels, population densities, and industrial structures across different provinces in China, data collection was conducted nationwide to ensure the sample's representativeness and generalizability.

In summary, this study draws upon independent variable data sourced from the National Bureau of Statistics of China, while dependent variable data undergoes calculation via the national air quality calculation platform. To ensure inclusivity, air quality data spanning from 2013 to 2019 and covering all provinces in China were gathered. This meticulous data collection process guarantees the reliability, authority, and representative nature of the dataset, forming a robust foundation for our research.

3.1.2 Data Cleaning and Outlier Handling

Data cleaning and exception handling are essential prerequisites before applying the multinomial logistic regression method. This crucial step serves to ensure data quality, reliability, and accuracy, thereby enhancing the precision and interpretability of the model.

The initial task in data cleaning involves addressing missing values, which can disrupt the model's fitting and inference process. Effective strategies must be employed to handle such instances. Common approaches include removing observations or variables with missing values, employing imputation techniques based on available information, or utilizing specialized methodologies like multiple imputation.

Another significant aspect of data cleaning pertains to outlier management. Outliers are extreme values that significantly deviate from the rest of the observations. These anomalies can have adverse effects on the model, leading to biased parameter estimates and distorted inference outcomes. Consequently, it becomes necessary to detect and appropriately address outliers. Established methods for this purpose include outlier detection based on statistical rules, normalization techniques employing boxplots or Z-scores, as well as advanced algorithms like Isolation Forest and LOF.

In addition to addressing missing values and outliers, data cleaning encompasses several other facets, such as handling duplicate values, performing variable transformation and standardization, and verifying/correcting data types. These steps contribute to maintaining data consistency, comparability, and suitability for analysis.

By effectively managing missing values, outliers, and other data issues, researchers can enhance the reliability and precision of their models, facilitating more accurate inferences and conclusions. It is important to note that these data cleaning and exception handling strategies should be tailored according to the specific characteristics of the dataset and align with both statistical principles and domain knowledge, thus minimizing biases and errors.

3.2 Methodology

AIC and AIC Criterion

The Akaike Information Criterion (AIC) is a widely used statistical tool for model selection, developed by Hirotsugu Akaike in the 1970s. It has gained significant popularity due to its ability to balance model fit and complexity, making it suitable for academic research across various scientific disciplines.

AIC is derived from information theory and is based on the principle of minimizing the information loss when approximating an unknown true data-generating process.

It provides a quantitative measure to compare and evaluate different models based on their goodness of fit. By considering both model quality and simplicity, AIC offers a way to choose the most appropriate model among a set of competing alternatives.

The AIC criterion can be mathematically represented as:

$$AIC = 2k - 2\ln(L) \tag{3.1}$$

where AIC represents the Akaike Information Criterion, k denotes the number of estimated parameters in the model, and L represents the maximized value of the likelihood function associated with the model.

The first component of the AIC formula, $2k$, penalizes models with more parameters. This penalty discourages overfitting, where a model becomes too complex and starts fitting random fluctuations in the data rather than the underlying patterns. The inclusion of $2k$ in the criterion ensures that simpler models are favored unless the increase in complexity significantly improves the fit to the data.

The second component of the AIC formula, $2ln(L)$, measures the goodness of fit of the model. The likelihood function, L, quantifies how well the model predicts the observed data. As the likelihood increases, the term $2ln(L)$ decreases, indicating a better fit. Models that closely match the data will have higher likelihoods, resulting in lower AIC values.

To select the best model using AIC, researchers compare the AIC values of different models fitted to the same dataset. The model with the lowest AIC is considered the most suitable for explaining the data. This model achieves a good balance between accuracy and complexity, providing a robust representation of the underlying processes.

AIC has become an essential tool in academic research, as it offers a rigorous and objective framework for model selection. By guiding researchers in choosing the most appropriate model, AIC promotes parsimony and ensures that the chosen model is not overly complex, reducing the risk of overfitting and improving generalizability. Its widespread use across various disciplines, including statistics, econometrics, ecology, and social sciences, demonstrates its versatility and usefulness in scientific investigations.

In conclusion, the Akaike Information Criterion (AIC) provides scientists and researchers with a statistical measure to select the best-fitting model while considering its complexity. With its mathematical formulation and principles rooted in information theory, AIC offers a robust and objective approach to model selection. By balancing goodness of fit and model simplicity, AIC enhances the credibility and reliability of academic research, facilitating the identification of models that effectively capture the underlying patterns in observed data.

3.3 Empirical result and explanations

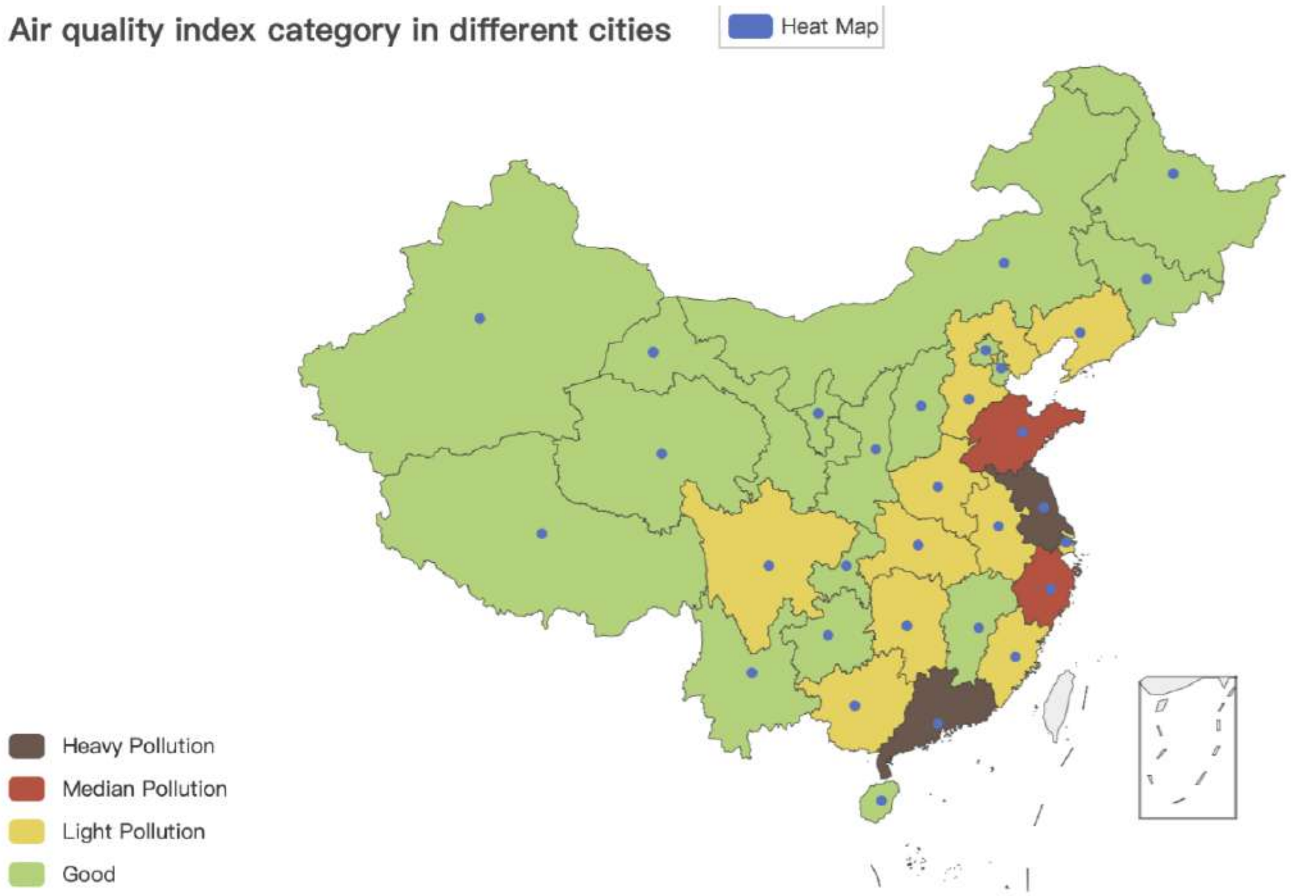


Figure 3.1: 31 capital cities air quality index situation at 2011

The provision of seven pictures depicting pollution conditions in various areas of China from 2011 to 2017 offers valuable insights into the trends and changes in air quality over time. The analysis of these visual representations allows for an academic and logical examination of the data, revealing patterns and drawing conclusions about the state of environmental protection efforts in China.

Firstly, it is noteworthy that half of the areas showcased good air quality throughout the years, as indicated by the green color blocks on the map. This observation

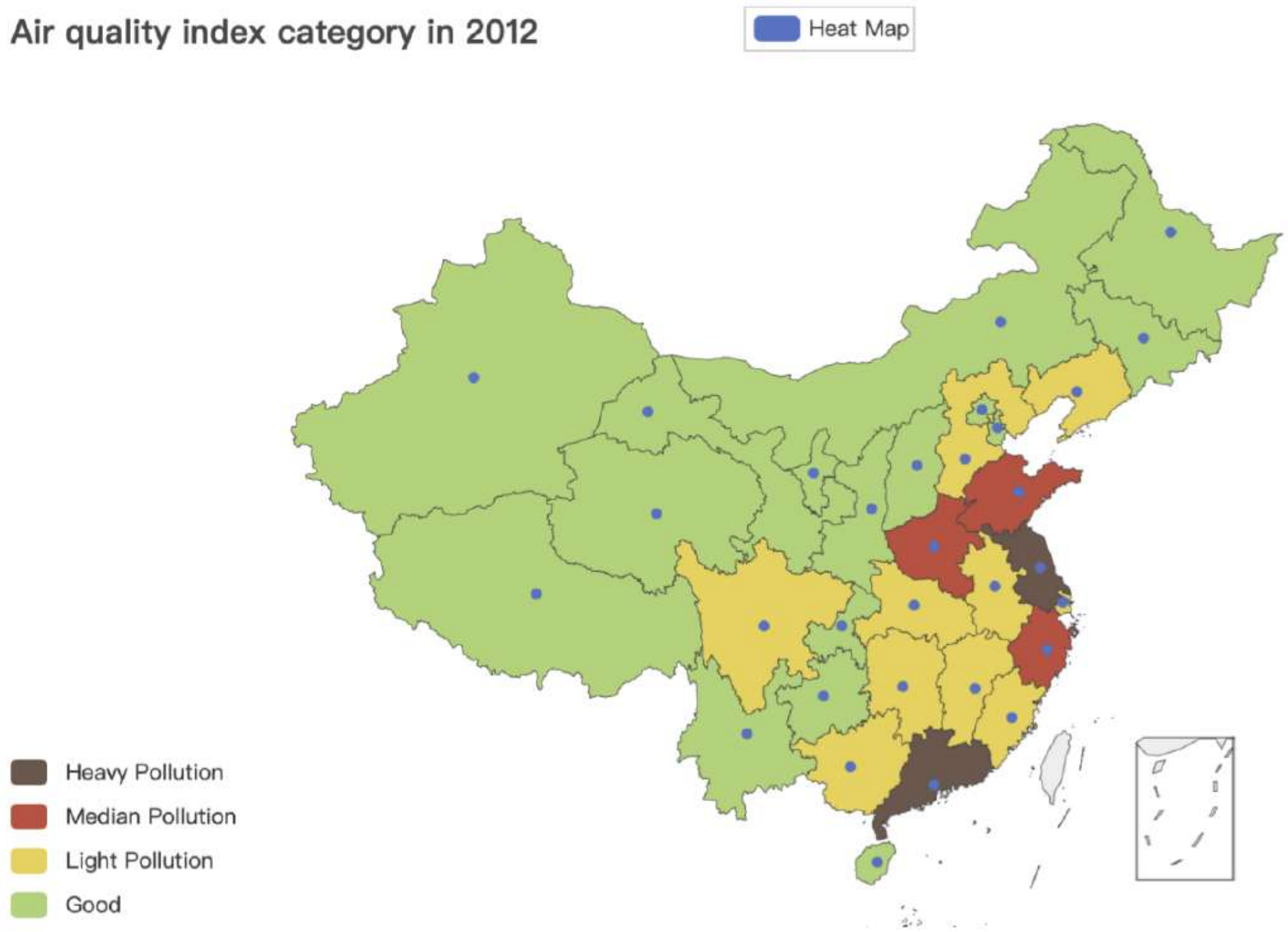


Figure 3.2: 31 capital cities air quality index situation at 2012

suggests that a significant portion of China's regions, cities, and provinces have been successful in maintaining or improving their air quality levels over the studied period. This finding aligns with the notion that more attention is being paid to environmental protection, leading to positive outcomes in terms of air quality.

Secondly, the presence of yellow color blocks, representing light pollution, indicates areas where air quality has not reached an optimal level but remains within a moderate range. Although less desirable than green, the prevalence of yellow suggests that efforts have been made to mitigate pollution and improve air quality in these regions. This demonstrates a positive trend towards addressing environmental concerns and taking steps to reduce pollution levels.

Furthermore, the occurrence of red color blocks, signifying moderate pollution, implies areas where the air quality falls below desired standards but still remains manageable. The presence of only two areas with severe pollution in the earlier years (2011-2016) and just one area with severe pollution in 2017 highlights a decline in the number of heavily polluted regions over time. This reduction can be

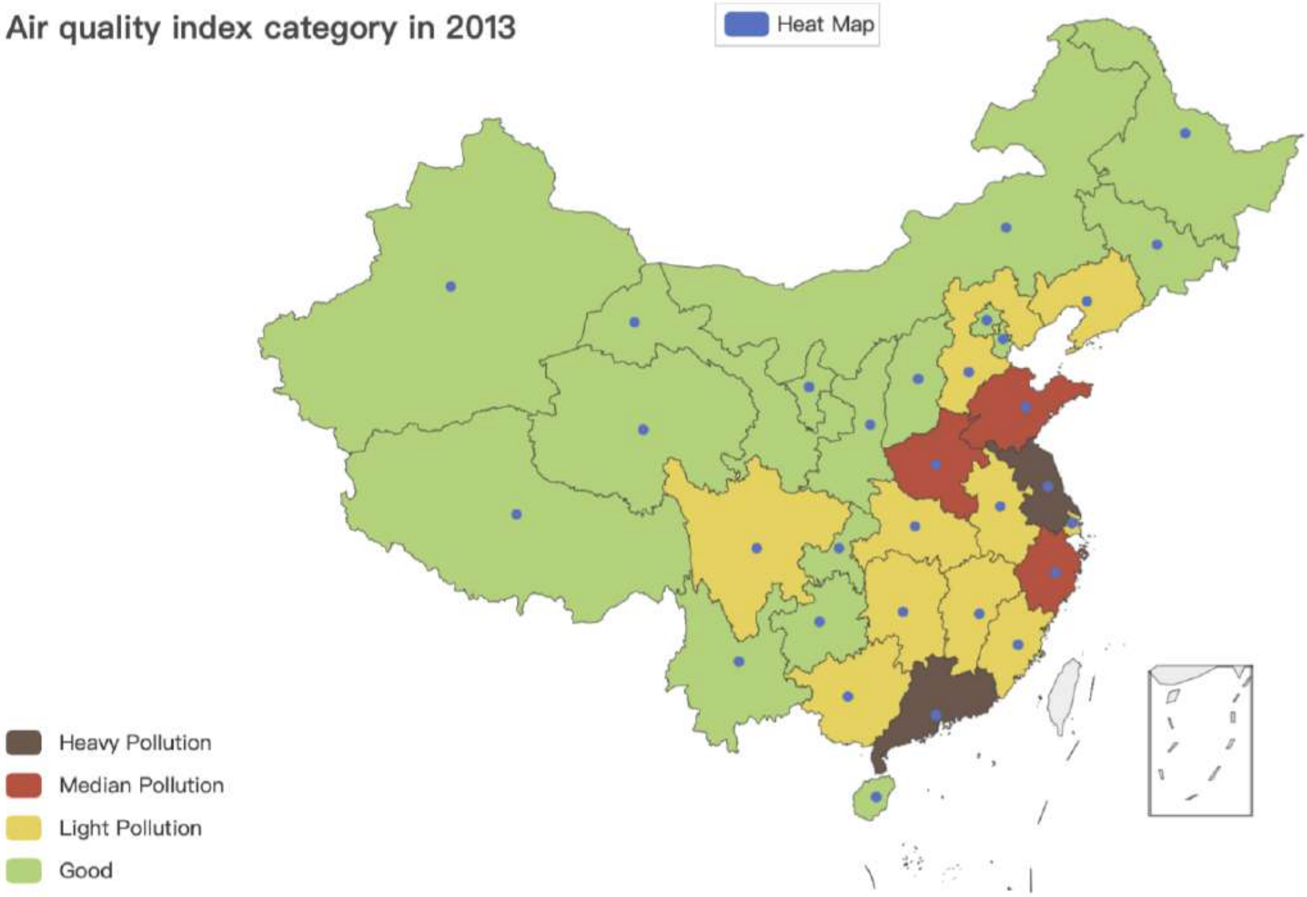


Figure 3.3: 31 capital cities air quality index situation at 2013

attributed to enhanced environmental protection measures and increased awareness of the detrimental effects of pollution, prompting authorities and communities to take action.

The consistent decrease in areas with severe pollution indicates progress in combating environmental degradation and underscores the effectiveness of environmental policies and initiatives. It suggests that the efforts invested in pollution control and environmental protection are yielding positive results, resulting in observable improvements in air quality across China.

From an academic perspective, the analysis of these pictures and the trends observed can contribute to a broader understanding of air pollution dynamics in China. It provides empirical evidence to support existing research on environmental policies, their implementation, and their impact on improving air quality. Additionally, it highlights the importance of continued efforts towards environmental protection and sustainable development.

Logically, the findings from this analysis suggest that as time progresses, more at-

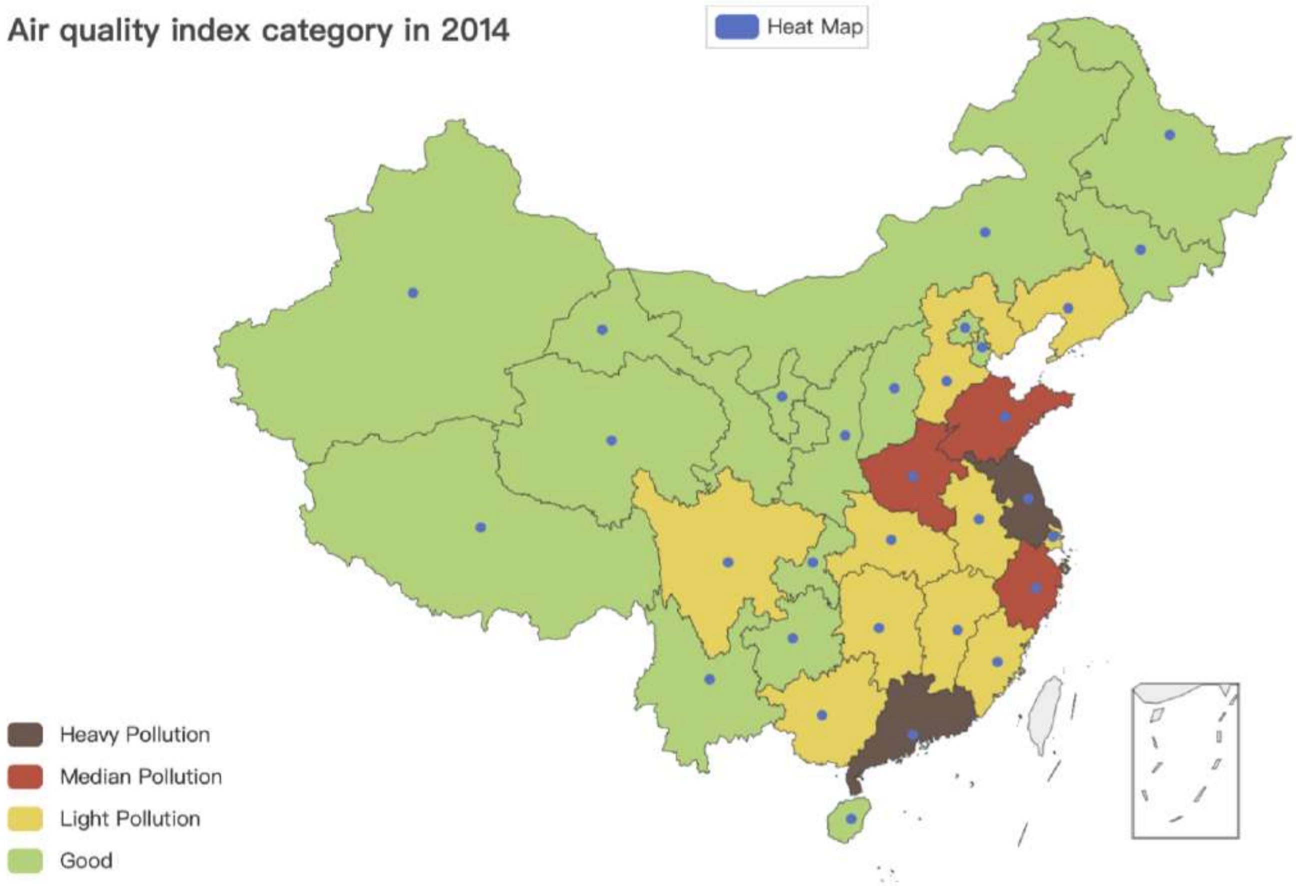


Figure 3.4: 31 capital cities air quality index situation at 2014

tention is being devoted to environmental protection in China. The decreasing areas with severe pollution indicate a growing awareness of the detrimental consequences of pollution and a shift towards adopting cleaner practices and technologies. This logical inference aligns with global trends emphasizing the significance of sustainable development and environmental stewardship.

The independent variables considered in this analysis include:

- Sulphur Dioxide (SO2)
- Nitrogen Dioxide (NO2)
- Particulate Matter with diameter less than 10 micrometers (PM10)
- Carbon Monoxide (CO)
- Ozone (O3)
- Particulate Matter with diameter less than 2.5 micrometers (PM2.5)
- Temperature

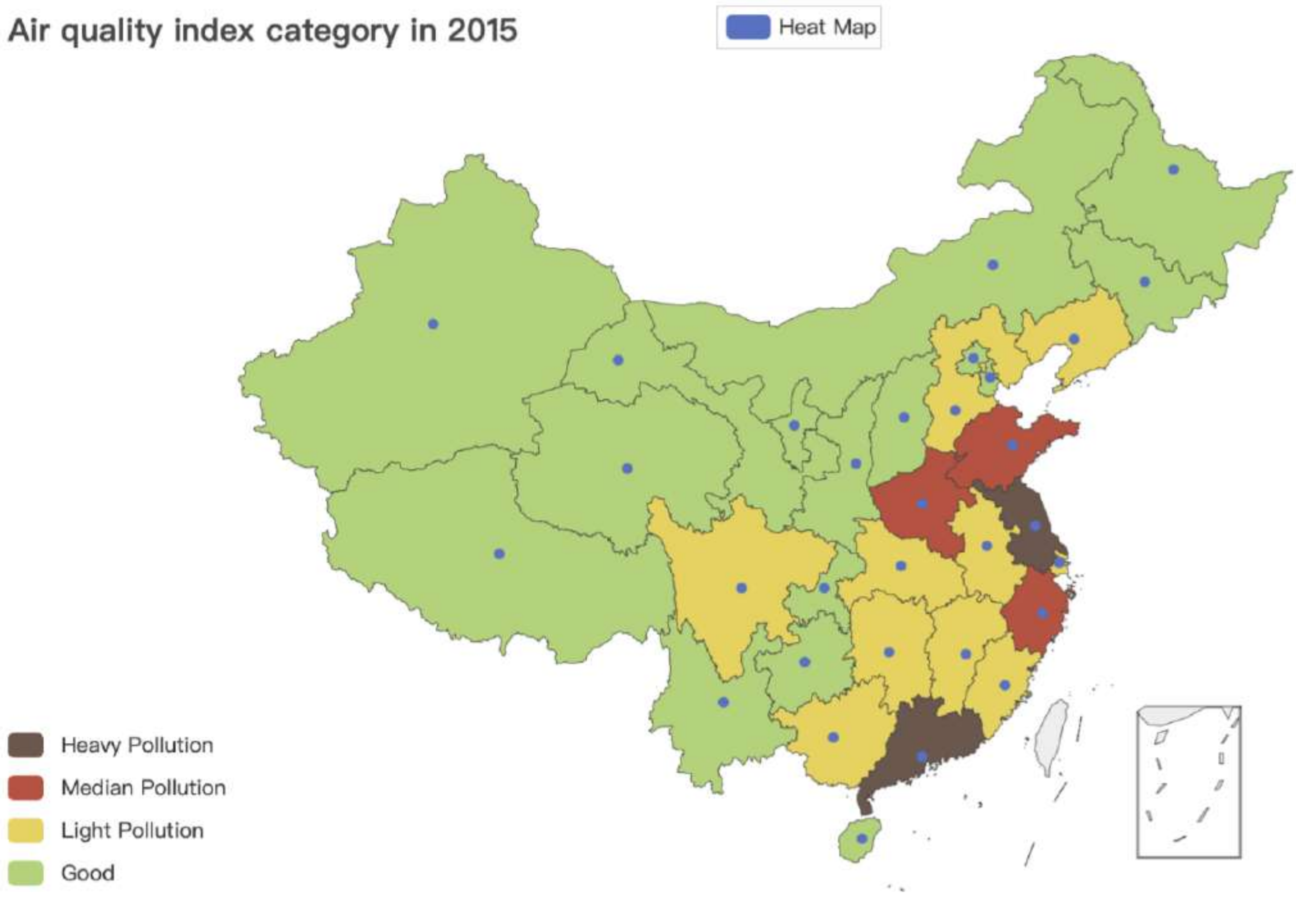


Figure 3.5: 31 capital cities air quality index situation at 2015

- Humidity
- Precipitation
- Sunshine

Furthermore, the dependent variable under investigation is the Air Quality Index (AQI).

The calculation of the Air Quality Index (AQI) involves a set of standardized formulas for converting pollutant concentrations into an overall index value.

A commonly used equation for calculating AQI is as follows:

$$AQI = \left(\frac{I_{high} - I_{low}}{C_{high} - C_{low}} \right) \times (C - C_{low}) + I_{low}$$

Where:

AQI represents the calculated Air Quality Index. I_{high} and I_{low} are the index values corresponding to the upper and lower breakpoints of the AQI scale, respectively. C_{high} and C_{low} are the concentration levels associated with the upper and lower

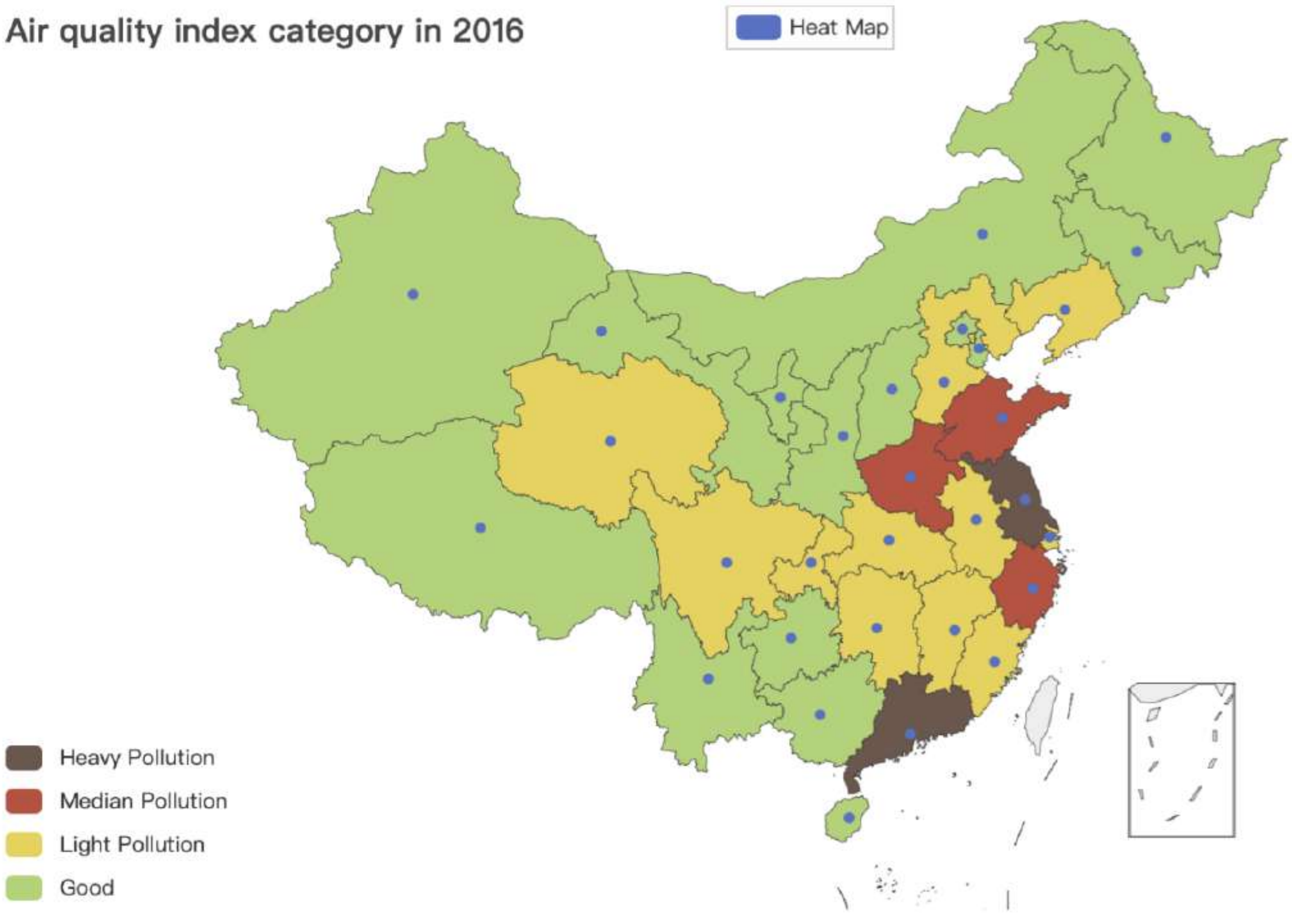


Figure 3.6: 31 capital cities air quality index situation at 2016

breakpoints, respectively. C represents the measured pollutant concentration. This formula linearly interpolates between the two breakpoints to estimate the AQI value based on the measured concentration level. The specific values of I_{high}, I_{low}, C_{high}, and C_{low} are determined by the respective air quality management agencies, and they vary depending on the pollutant and air quality category thresholds defined for that particular region.

China Air Quality Index Model Applied In 2017

The linear regression analysis was conducted utilizing the Ordinary Least Squares (OLS) method in Python, aiming to examine the relationship between the dependent variable, ln(AQI), and the explanatory variables encompassing pollutant items and meteorological factors. The model employed a constant term to account for any inherent bias in the data.

Upon executing the regression analysis, the estimated parameters and the results of significance tests were obtained. The collected data are summarized in the ac-

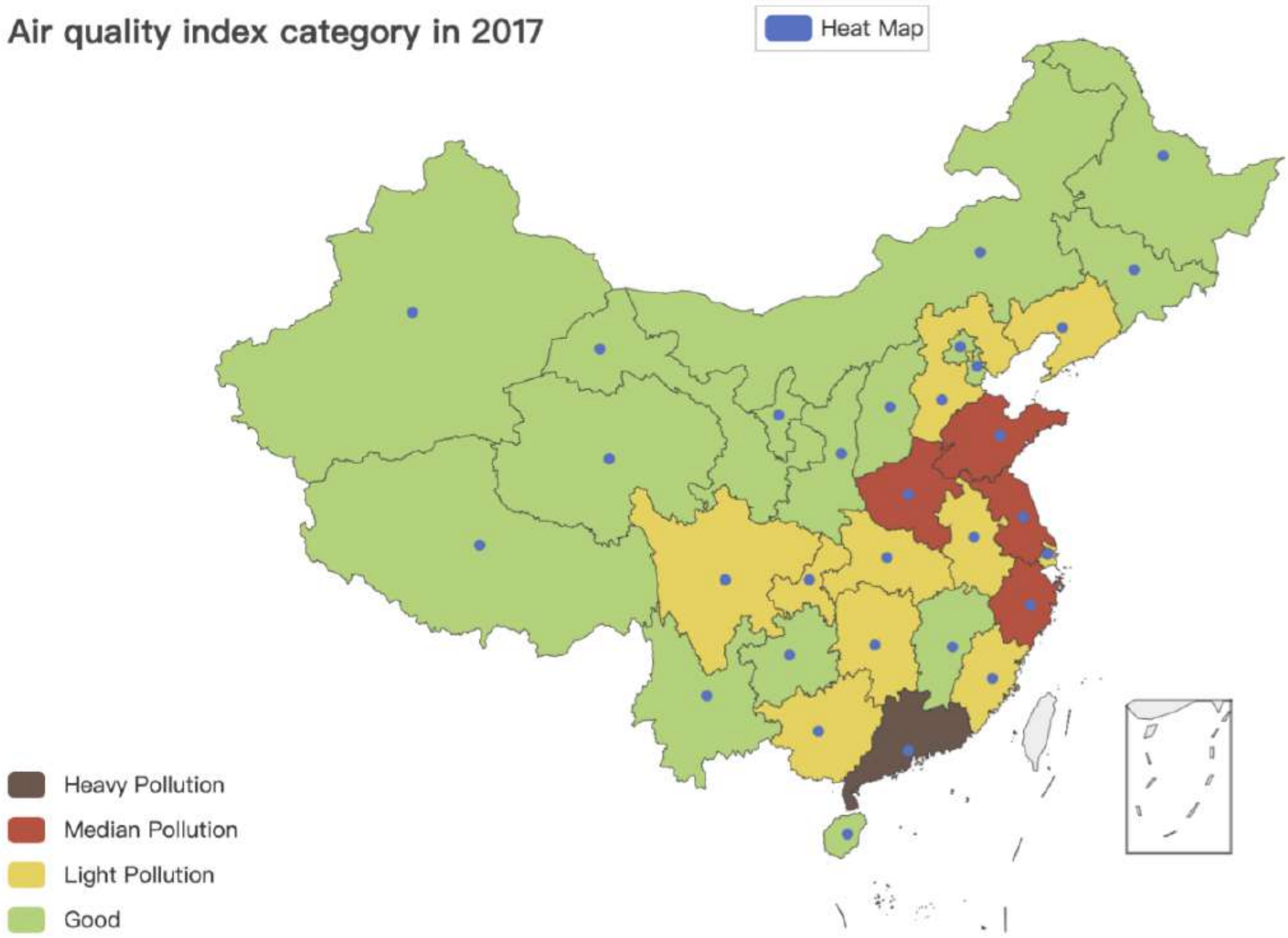


Figure 3.7: 31 capital cities air quality index situation at 2017

companying Table 3.1. By analyzing the statistical significance of the estimated coefficients, insights into the influence of the independent variables on the natural logarithm of AQI can be derived.

The p-value is a measure of statistical significance and assesses the likelihood that the observed relationship between each independent variable and the air quality index is due to chance.

The p-value for each variable indicates the probability of observing a coefficient as extreme or more extreme than the one estimated, assuming the null hypothesis that there is no relationship between the independent variable and the air quality index. If the p-value is below a predetermined significance level (typically 0.05), it suggests that the variable has a statistically significant association with the air quality index.

In this analysis, the variable PM10 demonstrates a significant association with the air quality index ($p = 0.003$). This implies that an increase in PM10 is associated with a higher air quality index. Other variables such as SO2, NO2, CO, O3, PM2.5, temperature, humidity, precipitation, and sunshine do not exhibit significant associations with the air quality index.

Table 3.1: OLS Regression Results

Variable	Coefficient	Std. Error	t-value	P-value	[0.025	0.975]
const	11,650	22,000	0.529	0.603(-)	-34,300	57,600
SO2	-4,771	2,974	-1.604	0.124(-)	-11,000	1,432
NO2	25,490	28,900	0.883	0.388(-)	-34,700	85,700
PM10	42,190	12,300	3.441	0.003(**)	16,600	67,800
CO	25,790	157,000	0.165	0.871(-)	-301,000	353,000
O3	-258.40	128.21	-2.015	0.057(.)	-525.84	9.05
PM2.5	795.37	1,333.77	0.596	0.558 (-)	-1,986.83	3,577.57
temp	-25.43	17.90	-1.421	0.171(-)	-62.76	11.90
humidity	831.93	541.29	1.537	0.140(-)	-297.19	1,961.05
precipitation	23.31	37.25	0.626	0.538(-)	-54.39	101.01
sunshine	-9.74	8.44	-1.154	0.262(-)	-27.34	7.87
Sig. codes	0 ’***’	0.001 ’ **’	0.01 ’*’	0.05 ’.’	0.1 ’-’	

Table 3.2: Regression Model Summary

Model Information				
Omnibus	Durbin-Watson	Prob(Omnibus)	Jarque-Bera (JB)	Skew
1.350	2.197	0.509	1.012	0.157
Prob(JB)	Kurtosis	Residual Normality Test p-value	Homoscedasticity Test p-value	
0.603	2.172	0.5091407387537017	0.50365152181532396	
Model Fit Statistics				
R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Log-Likelihood
0.948	0.921	36.14	1.30e-10	-373.75
No. Observations	AIC	Df Residuals	Df Model	BIC
31	769.5	20	10	785.3

According to the regression results in Table 3.1, we can get the regression model:

$$\begin{aligned}\hat{y}_{2017} =&11,650 - 4,771 \cdot x_1 + 25490 \cdot x2 + 42190 \cdot x_3 + 25790 \cdot x_4 - 258.4 \cdot x_5 + \\ &795.37 \cdot x_6 - 25.43 \cdot x_7 + 831.93 \cdot x_8 + 23.31 \cdot x_9 - 9.74 \cdot x_{10}\end{aligned} \tag{3.2}$$

The regression model summary presented in Table 3.2 provides valuable insights into the performance and goodness-of-fit of the model for predicting the air quality index.

The model demonstrates a strong fit to the data, as indicated by a high R-squared value of 0.948. This suggests that approximately 94.8% of the variability in the air quality index can be explained by the predictor variables included in the model. The adjusted R-squared value of 0.921 confirms that the model adequately accounts for the degrees of freedom used.

The F-statistic of 36.14 with a very low p-value of 1.30e-10 indicates that the overall regression is statistically significant. This implies that at least one of the predictor variables has a significant effect on the air quality index.

Moving on to the diagnostic tests, the Omnibus statistic tests the overall normality assumption of the residuals. With an Omnibus value of 1.350 and a corresponding p-value of 0.509, there is no significant evidence to suggest that the residuals deviate significantly from a normal distribution. Therefore, the assumption of normality holds reasonably well.

The Durbin-Watson statistic examines the presence of autocorrelation in the residuals. In this case, a value of 2.197 suggests no significant autocorrelation issues. This implies that the residuals are independent and do not exhibit systematic patterns.

The Jarque-Bera (JB) statistic and skewness measure the skewness and kurtosis of the residuals. With a JB value of 1.012 and a skewness of 0.157, there is no substantial evidence of departure from normal distribution assumptions. Additionally, the p-value of 0.603 further supports the adequacy of the normality assumption.

The homoscedasticity test assesses whether the residuals have constant variance across the range of predictor variables. The p-value of 0.503 indicates that there is no significant violation of the homoscedasticity assumption, suggesting that the model's residuals exhibit a relatively constant variance.

In summary, the diagnostic tests indicate that the model adequately meets crucial assumptions in regression analysis, including normality of residuals, absence of autocorrelation, and constant variance.

According to the OLS regression results shown in Table 3.1, the variable "CO" exhibits the highest p-value of 0.871 among all predictor variables. In line with the stepwise regression method, which aims to select the most influential variables, it is logical to eliminate the variable "CO" from the model first.

This decision aligns with the principle of feature selection in regression analysis, where variables deemed to have little or no impact on the response variable are progressively removed. By eliminating the variable "CO" with a high p-value, we reduce the complexity of the model and focus on more significant predictors that exhibit stronger relationships with the air quality index.

To ensure the significance of all variables, we employed the backward elimination

Table 3.3: Dynamics of Quality Indicators during Explanatory Variable Selection

Variable	R^2	R^2_{adj}	F-value	p-value(F)	AIC	p(RNT)	p(HT)	DW
-CO	0.947	0.925	42.10	1.9391×10^{-11}	767.6	0.5864	0.5134	2.203
-precip.	0.947	0.927	48.68	3.1992×10^{-12}	766.1	0.6111	0.6635	2.217
-PM2.5	0.946	0.930	57.54	4.4795×10^{-13}	764.4	0.6911	0.7104	2.198
-NO2	0.945	0.931	68.62	6.3971×10^{-14}	763.0	0.9670	0.7930	2.204
-temp	0.941	0.929	79.71	1.5308×10^{-14}	763.2	0.9473	0.7216	2.328
-hum.	0.941	0.931	102.9	1.5139×10^{-15}	761.4	0.8427	0.7183	2.384

stepwise regression method. During each step, we systematically removed the least significant variable based on the Student's t-test. Our aim was to maximize the F-statistic value while maintaining statistical significance.

In Table 3.3, we observed a consistent increase in the F-statistic value with each step. By eliminating the variable associated with the highest p-value from the Student's t-test, we were able to achieve the maximum F-statistic value. The control variables, which had relatively lower levels of significance, were eliminated at each stage according to the defined criterion.

Throughout the analysis, we conducted tests for Autocorrelation, Residual Normality, and Homoscedasticity at each step. These diagnostic tests allowed us to assess the assumptions of the model and evaluate its validity.

The resulting final model for this year is as follows:

$$\hat{y}_{2017} = -203.1495 - -2791.0443 \cdot x_1 + 52210 \cdot x_3 - 198.3622 \cdot x_5 - 16.0762 \cdot x_{10} \quad (3.3)$$

By employing the backward elimination stepwise regression method and conducting comprehensive diagnostic tests, we ensured that all variables included in the final model were statistically significant.

Based on the table, a logical conclusion can be drawn that the model resulting from the stepwise regression method, which aims to select the most relevant explanatory variables, coincides with the final model by having the smallest AIC value. The Akaike Information Criterion (AIC) is a widely accepted measure for assessing the trade-off between model goodness of fit and complexity. Thus, identifying the model with the minimum AIC value through stepwise regression signifies its superior performance in achieving the optimal balance between these factors. This finding holds academic significance as it highlights the effectiveness of the stepwise regression approach in arriving at the final model with the most favorable AIC value.

China Air Quality Index Model Applied In 2011-2016

Table 3.4: The model result after stepwise regression at 2011-2016

Time	Elimination Variable	R^2	R^2_{adj}	F-value	p-value(F)	AIC	p(RNT)	p(HT)	DW
2011	**3,4,5,6,10,8,7**	0.913	0.904	94.69	1.91×10^{-14}	767.3	0.2400	0.5812	1.561
2012	**7,10,5,6,3,4,8**	0.908	0.898	89.02	4.07×10^{-14}	771.5	0.2384	0.3255	1.393
2013	**6,9,8,10,3,5,4**	0.913	0.904	57.54	1.87×10^{-13}	770.9	0.5382	0.5705	1.586
2014	**6,4,5,8,10,3,7**	0.931	0.923	121.2	8.89×10^{-16}	767.0	0.1173	0.2471	1.864
2015	**6,7,5,8,3,4,10**	0.905	0.895	86.06	6.15×10^{-14}	777.3	0.5218	0.5907	1.855
2016	**5,10,7,8,4,6,9**	0.869	0.855	59.92	4.63×10^{-12}	785.7	0.4287	0.3906	1.885

Table 3.4 presents the results of stepwise regression conducted for each year, resulting in the identification of the final models. The "Elimination Variable" column indicates the order in which variables were eliminated during the stepwise regression process based on their lack of significance.

Firstly, examining the R^2 and R^2_{adj} values, it can be observed that the final models exhibit high goodness of fit across all years, ranging from 0.869 to 0.931. This indicates that a substantial proportion of the variations in the dependent variable can be explained by the selected set of independent variables.

Secondly, the F-value and its associated p-value (F) provide evidence supporting the overall statistical significance of the final models. Notably, the F-values range from 57.54 to 121.2, with corresponding p-values (F) ranging from 8.89×10^{-16} to 4.63×10^{-12}. These small p-values suggest that the final models are highly statistically significant, further reinforcing their adequacy in explaining the response variable.

The Akaike Information Criterion (AIC) is an essential criterion for model selection, balancing goodness of fit and model complexity. In this regard, it is noteworthy that the AIC values of the final models range from 767.0 to 785.7. The model with the smallest AIC value, i.e., 767.0, is obtained for the year 2014. This finding suggests that the model based on the elimination of variables in a specific order effectively achieves the optimal trade-off between model performance and parsimony for that particular year.

Additionally, examining the Durbin-Watson (DW) statistic provides insights into the presence of autocorrelation within the models. The DW values range from

1.393 to 1.885, indicating that there is no significant autocorrelation present in the residuals of the final models.

In summary, the analysis of Table 3.4 reveals that the stepwise regression technique consistently identifies final models with high goodness of fit and statistical significance across multiple years. By considering the AIC values, the elimination variable sequence, and the absence of autocorrelation, these models provide a robust representation of the relationship between the dependent and independent variables.

The final models obtained through stepwise regression for each year reveal that all variables included in the models are statistically significant. These models exhibit varying forms across different years, indicating the presence of unique relationships between the dependent and independent variables within each timeframe.

In 2011 year the final model (AIC = 767.3):

$$\hat{y_{2011}} = -1.471 \cdot 10^4 - 1856.64 \cdot x_1 + 4.728 \cdot 10^4 \cdot x_2 - 17.04 \cdot x_9 \tag{3.4}$$

In 2011 the model exhibiting the lowest Akaike Information Criterion (AIC = 767) :

$$\hat{y_{2011}} = -1.924 \cdot 10^4 - 1746.56 \cdot x_1 + 4.609 \cdot 10^4 \cdot x_2 + 3.693 \cdot x_7 - 28.199 \cdot x_9 \tag{3.5}$$

In 2012 year the final model(AIC = 771.5):

$$\hat{y_{2012}} = -2.068 \cdot 10^4 - 1740.56 \cdot x_1 + 4.728 \cdot 10^4 \cdot x_2 - 17.63 \cdot x_9 \tag{3.6}$$

In 2012 the model exhibiting the lowest Akaike Information Criterion (AIC = 769.7) :

$$\hat{y_{2012}} = -1.673 \cdot 10^4 - 1891.88 \cdot x_1 + 5.052 \cdot 10^4 \cdot x_2 - 447.746 \cdot x_8 - 11.456 \cdot x_9 \tag{3.7}$$

In 2013 the final model, at the same time exhibiting the lowest Akaike Information Criterion (AIC = 770.9):

$$\hat{y_{2013}} = -1.519 \cdot 10^4 - 1855.32 \cdot x_1 + 5.101 \cdot 10^4 \cdot x_2 - 9.734 \cdot x_7 \tag{3.8}$$

In 2014 year the final model, at the same time exhibiting the lowest Akaike

Information Criterion (AIC = 767):

$$\hat{y_{2014}} = -3.82 \cdot 10^4 - 1746.39 \cdot x_1 + 5.431 \cdot 10^4 \cdot x_2 - 42.678 \cdot x_9 \tag{3.9}$$

In 2015 year the final model, at the same time exhibiting the lowest Akaike Information Criterion (AIC = 777.3):

$$\hat{y_{2015}} = -1.959 \cdot 10^4 - 1706.58 \cdot x_1 + 5.473 \cdot 10^4 \cdot x_2 - 52.269 \cdot x_9 \tag{3.10}$$

In 2016 year the final model, at the same time exhibiting the lowest Akaike Information Criterion (AIC = 785.7):

$$\hat{y_{2016}} = 1.925 \cdot 10^4 - 6092.73 \cdot x_1 + 1.034 \cdot 10^5 \cdot x_2 - 7377.13 \cdot x_3 \tag{3.11}$$

Empirical result

Table 3.5: Final Model Results from 2011 to 2017

Time	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
2011	*	*	-	-	-	-	-	-	*	-
2012	*	*	-	-	-	-	-	*	*	-
2013	*	*	-	-	-	-	*	-	-	-
2014	*	*	-	-	-	-	-	-	*	-
2015	*	*	-	-	-	-	-	-	*	-
2016	*	*	*	-	-	-	-	-	-	-
2017	*	-	*	-	*	-	-	-	-	*
Count	7	6	2	0	1	0	1	1	4	1

In the statistical analysis of China's air quality index, a multiple regression model was constructed using the stepwise elimination method based on data from 2011 to 2017. The model aimed to investigate the relationship between the air quality index, which served as the target variable, and various environmental factors, namely: SO2 (x_1), NO2 (x_2), PM10 (x_3), CO (x_4), O3 (x_5), PM2.5 (x_6), Temperature (x_7), Humidity (x_8), Precipitation (x_9), and Sunshine (x_{10}).

The table labeled "Final Model Results from 2011 to 2017" provides valuable insights into the regression models developed for each year. Each row represents a specific year, while the columns correspond to the aforementioned variables. The presence of an asterisk (*) in a cell indicates the inclusion of that variable in the final regression model for that particular year. The "Count" row at the bottom of the table represents the number of times each factor was included in the regressors of the final models constructed from 2011 to 2017.

Analyzing the table, it is evident that factor x_1 (SO2) appears in all models throughout the years, showcasing its consistent impact on the air quality index from 2011 to 2017. Conversely, factors x_4 (CO) and x_6 (PM2.5) are consistently absent from the regression equations, indicating their relatively minimal influence on the air quality index during this period.

Furthermore, the variable x_2 (NO2) is frequently included in the regression models, suggesting its significance among the regressors. Similarly, factor x_9 (Precipitation) exhibits notable relevance in the final models.

3.4 Conclusion to the chapter

This chapter primarily employs the stepwise regression technique to construct models for the air quality index. When applying the stepwise regression method to analyze the air quality index data, it was observed that the selected model consistently differed from the one with the lowest AIC, except for the years 2011 and 2012.

Through the experiment, it is evident that factor x_1 (SO2) appears in all models throughout the years, showcasing its consistent impact on the air quality index from 2011 to 2017. Conversely, factors x_4 (CO) and x_6 (PM2.5) are consistently absent from the regression equations, indicating their relatively minimal influence on the air quality index during this period.

Furthermore, the variable x_2(NO2) is frequently included in the regression models, suggesting its significance among the regressors. Similarly, factor x9 (Precipitation) exhibits notable relevance in the final models.

Conduct a brief analysis based on the actual situation .The significant and sustained impact of SO2 on the air quality index from 2011 to 2017 can be attributed to several factors. SO2, derived mainly from coal combustion, industrial processes, and vehicle emissions, has been a major pollutant in China due to extensive coal usage and inadequate emission control measures in the past few decades. Long-term exposure to high levels of SO2 poses severe health risks and contributes to the formation of acid rain. Although the Chinese government has implemented various policies and measures to reduce SO2 emissions, the complex nature of emission sources and the cumulative effect require continuous efforts over an extended period to achieve significant improvements in air quality.

On the other hand, CO and PM2.5 demonstrate relatively minor impacts on the air quality index. CO, a colorless and odorless toxic gas primarily emitted from coal combustion, vehicle exhaust, and industrial processes, has a short atmospheric lifetime, resulting in rapid decrease in concentration with changing meteorological conditions. Similarly, PM2.5, consisting of fine particles suspended in the air, including dust, smoke, and vehicle emissions, exhibits varying concentrations influenced by diverse sources and meteorological conditions. While both CO and PM2.5 contribute to air pollution and adverse health effects, their influences may not be as stable or persistent as SO2.

Furthermore, NO2, a nitrogen dioxide produced from combustion processes such as coal burning, vehicle emissions, and industrial activities, demonstrates no significant impact on the air quality index in 2017. This could be due to additional factors or measures leading to reduced emissions or improved control efficiency during that specific period. Additionally, the concentration of NO2 is influenced by meteorological conditions, emission source locations, and local surroundings. As a result, certain regions or specific time periods may experience a lack of significant impact from NO2 on the air quality index.

In conclusion, the sustained and significant impact of SO2 on the air quality index is attributed to its widespread emission sources, long-term health risks, and challenges involved in the improvement process. Conversely, the relatively minor effects of CO and PM2.5 are associated with the complexity of their emission sources, variability in concentrations, and challenges in controlling emissions. The lack of significant impact from NO2 in 2017 may be due to other intervening factors, improved control efficiency, and regional or temporal variations.

Chapter 4

Neural network models for air quality evaluation system

Most of results presented in the chapter are published in paper [102].

4.1 Standardized usage procedures of machine learning

The standardized usage process of machine learning encompasses several essential steps, among which the segmentation of datasets and ensuring the consistency of training data are of utmost importance. This section provides a detailed exploration of the logical flow and individual steps involved in this process.

Dataset Segmentation: Before embarking on machine learning tasks, it is imperative to partition the available dataset into distinct subsets: the training set, validation set, and test set. Such division facilitates model performance evaluation and parameter tuning.

- Training Set: The training set constitutes a subset of data employed to train the model. Typically, it represents a substantial portion of the overall dataset, enabling the model to acquire appropriate patterns and regularities.

- Validation Set: The validation set, as another subset of data, is used for hyperparameter selection, model tuning, and performance evaluation. By assessing and comparing different models using the validation set, one can identify the optimal model and make necessary adjustments.

- Test Set: The test set serves as an independent subset of data utilized for the ultimate evaluation of model performance. It should remain distinct from the training and validation sets to provide an accurate assessment of the model's generalization capability. Assessing the model against the test set yields an accurate estimation of its predictive performance in real-world scenarios.

Ensuring Training Data Consistency: Maintaining the consistency of training data plays a pivotal role in machine learning tasks. The following key steps are involved:

- Feature Normalization: Due to varying units or scales, different features may necessitate normalization to ensure equitable treatment by the model. Common techniques include standardization (transforming data into a distribution with a mean of 0 and variance of 1) and normalization (scaling data within the [0, 1] range).

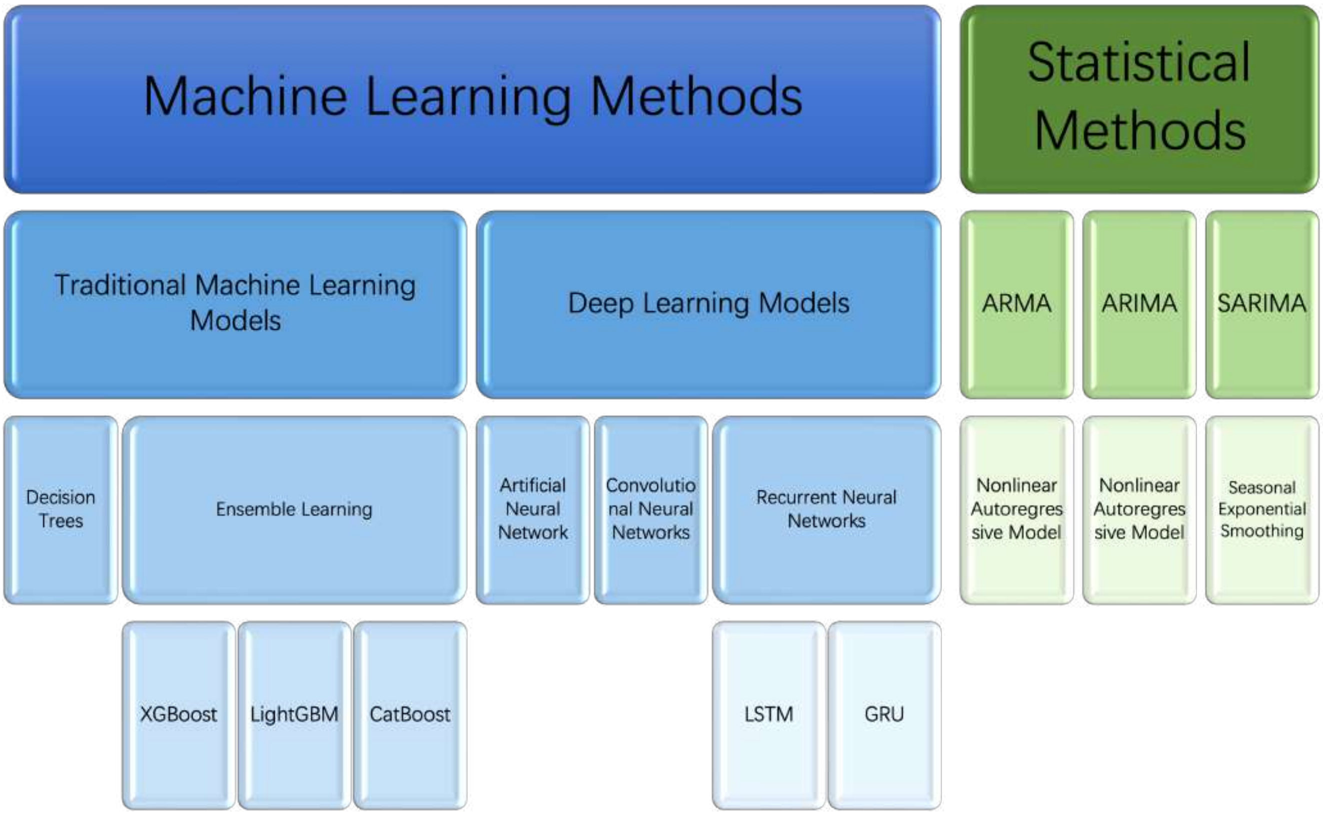
- Handling Missing Values: Thoroughly checking and addressing missing values in the training data is vital for successful model training and result accuracy. Strategies for filling in missing values often involve utilizing the feature's mean, median, or other suitable values.

- Dealing with Outliers: The identification and treatment of outliers within the training data are essential. Outliers can adversely affect model performance, and therefore, they should be identified and addressed using statistical methods or domain knowledge to ensure they do not disrupt the model training and prediction process.

By executing dataset segmentation and ensuring training data consistency, one can construct reliable and effective machine learning models. These steps enhance model performance, generalization capability, and stability, providing a robust foundation for further evaluation and fine-tuning. Adopting a standardized usage process ensures consistency and predictive power across different datasets, rendering machine learning models valuable decision support tools in practical applications.

ARMA (Autoregressive Moving Average), ARIMA (Autoregressive Integrated Moving Average), and SARIMA (Seasonal ARIMA) are statistical models commonly used for time series forecasting.

ARMA model combines autoregressive (AR) and moving average (MA) components to model time series data. The AR component predicts the current value based


Machine Learning Methods
Statistical Methods
Traditional Machine Learning Models
Deep Learning Models
ARMA
ARIMA
SARIMA
Decision Trees
Ensemble Learning
Artificial Neural Network
Convolutional Neural Networks
Recurrent Neural Networks
Nonlinear Autoregressive Model
Nonlinear Autoregressive Model
Seasonal Exponential Smoothing
XGBoost
LightGBM
CatBoost
LSTM
GRU

on past observations, assuming a linear relationship between the current value and previous values. The MA component predicts the current value based on the residual errors, assuming a linear relationship between the current value and past error terms.

ARIMA model extends the ARMA model by incorporating differencing operations to handle non-stationary time series. Differencing transforms the original time series into a stationary one by subtracting the previous observation from the current observation. This helps capture trends and seasonality in the data more accurately.

SARIMA model further extends ARIMA by introducing seasonal differencing and seasonal AR and MA terms. It is designed to handle time series data with clear seasonality patterns. Seasonal differencing removes the seasonal component, while seasonal AR and MA terms capture the seasonal dependencies in the data.

While ARMA, ARIMA, and SARIMA models have been widely used for time series forecasting, there are scenarios where machine learning models may be preferred in long-term time series forecasting.

Complex Patterns: Machine learning models, such as neural networks and deep learning models, can capture complex nonlinear relationships and patterns in the data that may not be well-suited for traditional statistical models like ARMA, ARIMA, and SARIMA. These models can learn intricate temporal dependencies and adapt to changing dynamics in the data.

Feature Extraction: Machine learning models can automatically extract relevant features from raw time series data, eliminating the need for manual feature engineering required in statistical models. This capability enables the models to uncover hidden patterns and extract meaningful representations from the data.

Scalability: In long-term time series forecasting, the volume and complexity of data may increase significantly. Machine learning models can handle large datasets efficiently, making them more suitable for scalability and handling high-dimensional data compared to traditional statistical models.

Flexibility: Machine learning models offer greater flexibility in modeling different types of time series, including those with nonlinear trends, non-Gaussian distributions, or irregular patterns. They can adapt to various data characteristics and accommodate different modeling assumptions.

Incorporating External Factors: Machine learning models can easily incorporate external factors or additional features that may influence the time series, allowing for enhanced predictive accuracy compared to traditional statistical models.

4.2 Data select

Data Source and Collection: The collection and sourcing of data play a vital role in the context of time series forecasting. In this particular scenario, the dataset originates from Kaggle, a widely recognized open data platform renowned for its vast array of datasets catering to research and analysis purposes.

For effective time series forecasting, the quality and diversity of the dataset are of utmost importance. In this case, the dataset covers a significant time span, ranging from March 1, 2013, to February 28, 2017, thereby encapsulating a considerable duration. Such employment of long-term data proves exceptionally advantageous for deep learning models, as they typically necessitate substantial amounts of data for training and generalization.

The PM2.5 measurements represent fine particulate matter with a diameter of 2.5 micrometers or smaller, which is particularly relevant to public health as it can penetrate the respiratory system and cause adverse effects. Atmospheric pressure, on the other hand, serves as an indicator of the movement of air masses, and its inclusion in the dataset can help researchers identify potential correlations between pressure changes and pollutant levels. Similarly, temperature and humidity are essential meteorological factors that may directly influence pollutant dispersion, chemical reactions, and atmospheric stability, impacting air quality levels accordingly. By considering all these diverse environmental parameters in the analysis, researchers can develop a more nuanced perspective on the dynamics of air quality. This comprehensive dataset facilitates the examination of the interplay between various factors and their combined effects on pollutant concentrations, ultimately enabling the creation of more accurate and reliable prediction models for effective air quality management.

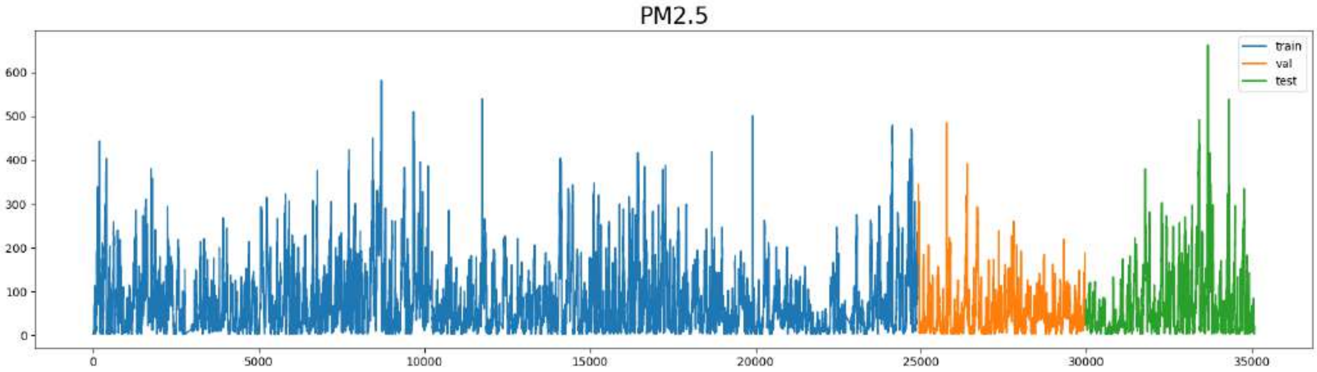
The magnitude of the dataset allows for the capture of a greater number of features and patterns, leading to heightened predictive capabilities within the model. Larger datasets provide increased opportunities for deep learning models to comprehend intricate temporal structures and dependencies present within the time series. Moreover, extensive datasets also mitigate the risk of overfitting and enhance the model's stability and robustness.

Given that Kaggle is an extensively utilized open data platform, the datasets available undergo rigorous scrutiny and validation, ensuring their reliability and reproducibility. This instills confidence among researchers and data scientists when engaging in data preprocessing and modeling procedures.

It should be considered that the utilization of large-scale datasets by deep learning models necessitates ample computational resources and time. Researchers must ensure adequate computing power and formulate reasonable schedules to accommodate the demands associated with training tasks when employing such datasets.

In this study, a specific time point serves as the demarcation between the training and test sets, rather than employing an arbitrary random selection process. This approach aligns more closely with the inherent logic of time series data. Approximately 70% of the elements are allocated to the training set. Considering the varying proportions of missing values across different types, we ensure an adequate number of features while forecasting Particulate Matter 2.5 (PM2.5) as the target variable in this research. Upon establishing a rational boundary, the proportions of the training set, test set, and validation set are determined to be 70%, 15%, and

15%, respectively.



Data Cleaning and Outlier Handling: Data preprocessing, including data cleaning and anomaly handling, is an essential phase when employing deep learning models for time series forecasting. These steps are instrumental in ensuring the accuracy, consistency, and reliability of the data, thereby enhancing the model's performance and robustness.

Data Cleaning: The initial step involves addressing issues such as missing values, duplicate entries, and outliers. Missing values can be filled using interpolation methods like linear or Lagrangian techniques. Duplicate entries can be directly eliminated. To tackle outliers, appropriate outlier detection techniques need to be employed.

Stabilization: Another crucial step involves transforming non-stationary time series into a stationary form. Non-stationary time series possess mean, variance, or covariance that vary with time, making it challenging for the model to capture meaningful patterns and regularities. Common stabilization methods include differencing (first or second order), logarithmic transformations, or other applicable techniques.

Anomaly Detection and Handling: Anomalies within the time series can significantly impact the model's predictive capability. Therefore, it is vital to employ anomaly detection techniques to identify and address these outliers. Machine learning methods, such as isolation forests or dedicated outlier detection algorithms, can be leveraged for detecting anomalies. Once identified, outliers can be managed through various approaches, such as removal, replacement, or correction based on the specific context.

Feature Engineering: Effective feature engineering plays a pivotal role in the success of deep learning models. It entails selecting and extracting relevant features from the time series data, enabling the model to capture intricate patterns and trends more effectively.

Data Standardization: To enhance the training effectiveness and convergence speed of deep learning models, data standardization becomes necessary. Normalization techniques, such as min-max normalization, can be applied to scale the data within a comparable range, mitigating issues arising from varying scales across different features.

In conclusion, conducting comprehensive data cleaning, anomaly handling, feature engineering, and data standardization are crucial pre-processing steps in deep learning models for time series forecasting. These procedures ensure the integrity and reliability of the data, ultimately contributing to accurate and robust predictions.

4.3 Basic Methodology

4.3.1 Artificial Neural Network(ANN)

When using Artificial Neural Network (ANN) for time series forecasting, a feed-forward neural network (FNN) structure can be employed. FNN consists of an input layer, hidden layers, and an output layer, where neurons in each layer are interconnected with weighted connections.

For time series forecasting task, we can use the historical observations of the time series as inputs, while the objective is to predict future values. The mathematical expressions for the ANN model are as follows:

Input to Hidden Layer:

$$h_1 = f(W_{in}x + b_{in})$$

Hidden Layer to Output Layer:

$$y = g(W_{out}h_1 + b_{out})$$

Here, x represents the input at the current time step, h_1 is the output of the hidden layer, W_{in} and b_{in} are the weight and bias terms from the input layer to the hidden layer, and W_{out} and b_{out} are the weight and bias terms from the hidden layer to the output layer. $f(\cdot)$ and $g(\cdot)$ denote activation functions.

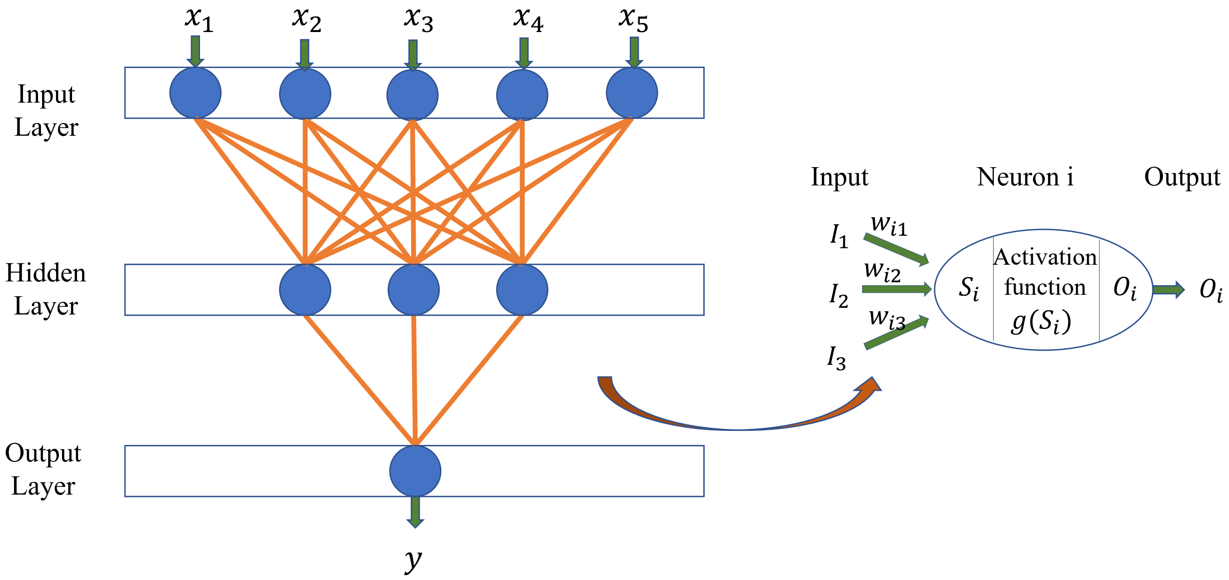


Figure 4.1: Network structure of ANN

In time series forecasting, regression tasks are commonly used, hence the output layer usually does not apply a non-linear transformation by using an activation function. Instead, it directly outputs the predicted value.

The training process of the ANN model typically involves minimizing a loss function. Commonly used loss functions include mean squared error (MSE), mean absolute error (MAE), etc. The model adjusts the connection weights through the backpropagation algorithm to minimize the loss and improve prediction performance.

In time series forecasting, it is important to consider the temporal correlation and sequential dependencies in the input data. A common approach is to use sliding window technique, taking the historical observations as input features and the next time step's observation as the target, to train the model for prediction.

By appropriately designing the network structure, selecting suitable activation and loss functions, the ANN model can achieve good performance in time series forecasting tasks. Techniques such as regularization, batch normalization, etc., can also be incorporated to enhance the model's generalization ability and stability.

4.3.2 Recurrent Neural Network(RNN)

RNN (Recurrent Neural Network) is a type of neural network architecture commonly used for time series forecasting tasks. It is particularly effective in capturing sequential dependencies and temporal patterns within the data. The key idea behind RNN is the introduction of recurrent connections, which allow information to be passed from previous time steps to current time steps.

The basic mathematical formulation of an RNN model for time series forecasting is as follows:

$$h_t = f(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

$$y_t = g(W_{hy}h_t + b_y)$$

where x_t represents the input at time step t, h_t denotes the hidden state at time step t, and y_t is the output at time step t. W_{hh}, W_{xh}, W_{hy}, b_h, and b_y are the weight matrices and bias vectors that need to be learned during training. $f(\cdot)$ and $g(\cdot)$ represent activation functions.

In the equations above, h_t is computed based on the input x_t and the hidden state from the previous time step h_{t-1}. This allows the model to learn and capture information from past observations that can influence the current prediction. The output y_t is then generated based on the current hidden state h_t.

During the training process, the parameters of the RNN model are optimized by minimizing a loss function that compares the predicted outputs with the ground truth values. The backpropagation through time (BPTT) algorithm is typically used to compute the gradients of the loss function with respect to the model parameters, enabling the update of the weights and biases.

It's important to note that in practice, additional variations of RNNs have been developed to overcome the vanishing/exploding gradient problem and improve learning long-term dependencies. Some popular variants include LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit). These variants introduce additional gating mechanisms that regulate the flow of information within the recurrent connections.

By leveraging the temporal dynamics captured by RNN models, it is possible

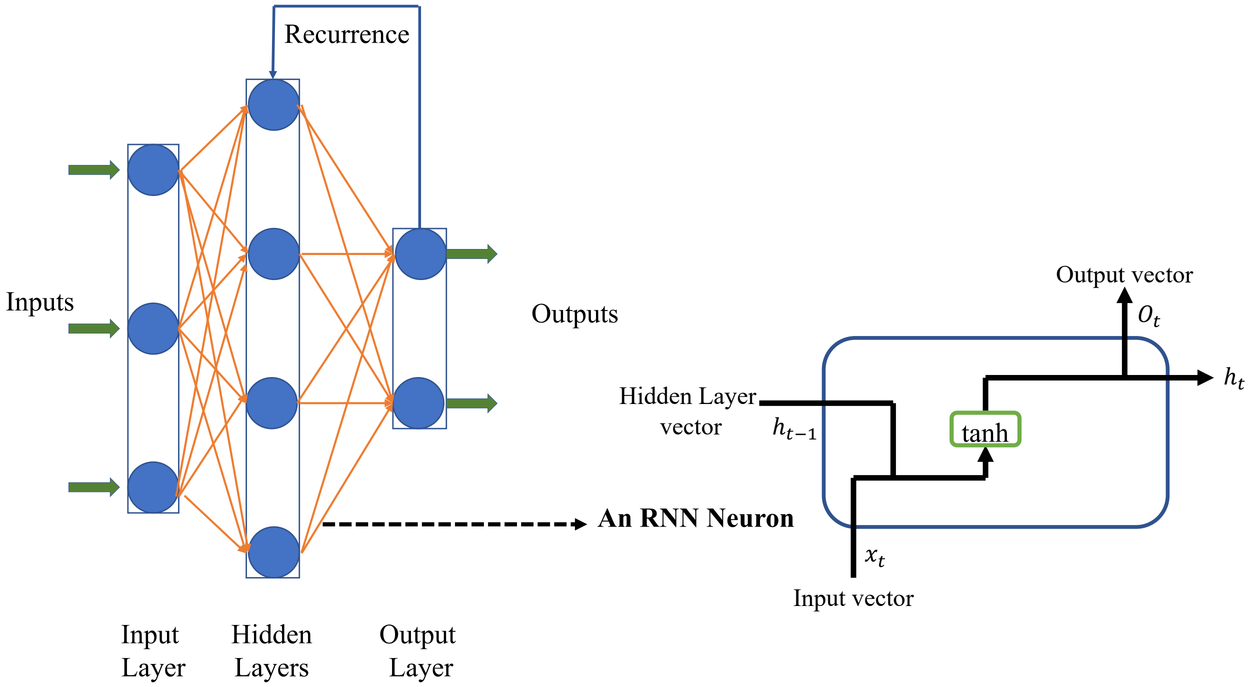


Figure 4.2: Network structure of traditional RNN

to effectively forecast future values in time series data. The choice of appropriate architectures and hyperparameters depends on the specific characteristics of the dataset and the forecasting task at hand.

4.3.3 Long Short-Term Memory(LSTM)

LSTM (Long Short-Term Memory) is a variant of Recurrent Neural Networks (RNNs) that has been widely used for time series forecasting tasks. LSTM models are specifically designed to address the problem of capturing long-term dependencies in sequential data.

The core component of an LSTM unit is the memory cell, which allows the network to retain and update information over multiple time steps. The mathematical formulation of an LSTM model for time series forecasting is as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$

$$h_t = o_t \odot \tanh(c_t)$$

$$y_t = g(W_{hy}h_t + b_y)$$

where x_t represents the input at time step t, h_t denotes the hidden state at time step t, c_t represents the cell state at time step t, and y_t is the output at time step t. W and b represent weight matrices and bias vectors that need to be learned during training. $\sigma(\cdot)$ denotes the sigmoid activation function and $(\odot)$ represents element-wise multiplication. f_t, i_t, and o_t represent the forget gate, input gate, and output gate respectively.

In the equations above, the LSTM model calculates three gates i_t, f_t, and o_t to control the flow of information in and out of the memory cell. The forget gate f_t determines which information to discard from the previous memory cell state c_{t-1}, based on the current input x_t and previous hidden state h_{t-1}. The input gate i_t controls which new information to store in the memory cell. The output gate o_t regulates the flow of information from the memory cell to the current hidden state. Finally, the hidden state h_t is obtained by applying the output gate to the cell state.

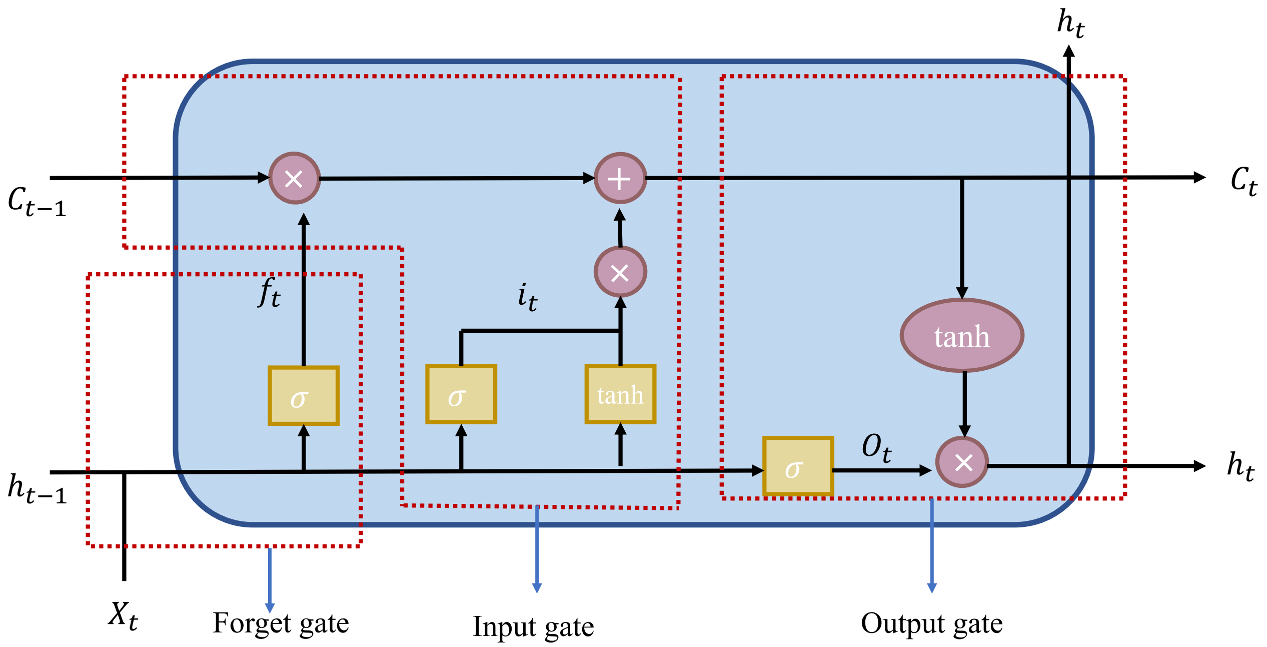


Figure 4.3: Network structure of LSTM

During training, the parameters of the LSTM model are optimized by minimizing a suitable loss function, typically using gradient-based optimization algorithms.

Backpropagation through time (BPTT) is commonly used to compute the gradients of the loss function with respect to the model parameters, allowing for the update of the weights and biases.

LSTM models have been proven effective in capturing long-term dependencies and handling vanishing/exploding gradient problems that can occur in traditional RNNs. By leveraging the memory cell, LSTM models can effectively capture and utilize relevant contextual information for accurate time series forecasting.

4.3.4 Gated Recurrent Unit(GRU)

GRU (Gated Recurrent Unit) is another variant of Recurrent Neural Networks (RNNs) that has gained popularity in time series forecasting tasks. GRU models are designed to capture long-term dependencies and address the vanishing/exploding gradient problem, similar to LSTM models.

The mathematical formulation of a GRU model for time series forecasting is as follows:

$$
\begin{aligned}
z_t &= \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z) \\
r_t &= \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r) \\
n_t &= \tanh(W_{xn}x_t + r_t \odot (W_{hn}h_{t-1}) + b_n) \\
h_t &= (1 - z_t) \odot n_t + z_t \odot h_{t-1} \\
y_t &= g(W_{hy}h_t + b_y)
\end{aligned}
$$

where x_t represents the input at time step t, h_t denotes the hidden state at time step t, and y_t is the output at time step t. W and b represent weight matrices and bias vectors that need to be learned during training. $\sigma(\cdot)$ denotes the sigmoid activation function and $(\odot)$ represents element-wise multiplication.

In the equations above, the GRU model introduces two gates: the update gate z_t and the reset gate r_t. The update gate controls how much of the previous hidden state h_{t-1} should be preserved and combined with the current hidden state candidate n_t. The reset gate determines how much of the previous hidden state h_{t-1} should be forgotten when computing the hidden state candidate n_t. Finally, the updated hidden state h_t is a combination of the previous hidden state weighted by the update gate and the current hidden state candidate weighted by $1 - z_t$.

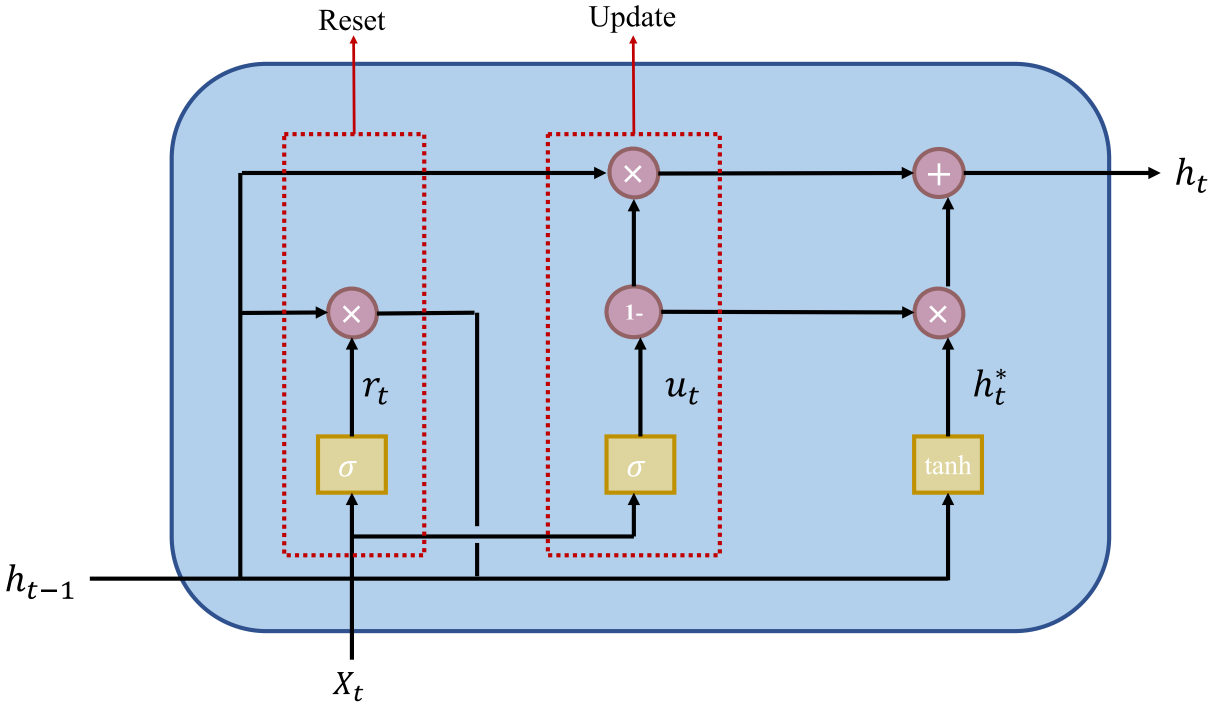


Figure 4.4: Network structure of GRU

During training, the parameters of the GRU model are optimized by minimizing a suitable loss function using gradient-based optimization algorithms. Backpropagation through time (BPTT) is commonly used to compute the gradients of the loss function with respect to the model parameters, allowing for the update of the weights and biases.

GRU models provide a simpler architecture compared to LSTM models while achieving similar performance in capturing long-term dependencies. They have been widely used in time series forecasting tasks and can effectively capture temporal patterns and dependencies in the data.

4.3.5 Bidirectional Recurrent Neural Network(Bi-RNN)

Bi-RNN (Bidirectional Recurrent Neural Network) is a variant of Recurrent Neural Networks (RNNs) that aims to capture both past and future information in a time series for improved forecasting performance. It combines two separate RNNs: one processing the sequence in a forward direction and the other in a backward direction.

The mathematical formulation of a Bi-RNN model for time series forecasting is as follows:

Forward RNN:

$$h_t^{\rightarrow} = f(W_{\rightarrow}x_t + U_{\rightarrow}h_{t-1}^{\rightarrow} + b_{\rightarrow})$$

Backward RNN:

$$h_t^{\leftarrow} = f(W_{\leftarrow}x_t + U_{\leftarrow}h_{t+1}^{\leftarrow} + b_{\leftarrow})$$

where x_t represents the input at time step t, $h_t^{\rightarrow}$ denotes the hidden state computed by the forward RNN, $h_t^{\leftarrow}$ represents the hidden state computed by the backward RNN, and $f(\cdot)$ represents the activation function.

In a Bi-RNN model, the forward RNN processes the sequence from the beginning to the end, while the backward RNN processes the sequence in the opposite direction. Each RNN maintains its own set of weights and biases $W_{\rightarrow}$, $U_{\rightarrow}$, $b_{\rightarrow}$ for the forward RNN and $W_{\leftarrow}$, $U_{\leftarrow}$, $b_{\leftarrow}$ for the backward RNN), which are learned during training.

To obtain the final hidden state at each time step, the forward and backward hidden states are concatenated:

$$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$$

where [; denotes concatenation.

The final output of the Bi-RNN model can be calculated as:

$$y_t = g(W_{hy}h_t + b_y)$$

where W_{hy} and b_y represent the weight matrix and bias vector for the output layer, and $g(\cdot)$ represents the activation function.

During training, the parameters of the Bi-RNN model are optimized by minimizing a suitable loss function using gradient-based optimization algorithms. The gradients are computed through backpropagation through time (BPTT) to update the weights and biases.

By considering both past and future information in the time series, Bi-RNN models can capture more comprehensive temporal dependencies and patterns, leading to improved forecasting performance compared to traditional RNNs. They have been successfully applied in various time series forecasting tasks.

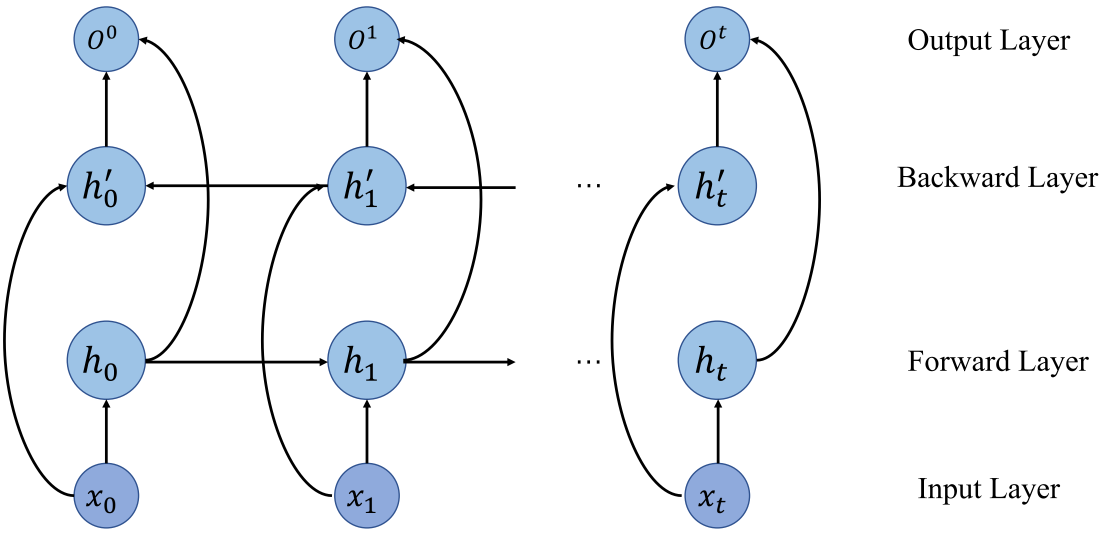


Figure 4.5: Network structure of Bi-RNN

Bidirectional Long Short-Term Memory(Bi-LSTM)

Bi-LSTM (Bidirectional Long Short-Term Memory) is a variant of Recurrent Neural Networks (RNNs) that combines the advantages of LSTM and Bidirectional RNNs to capture both past and future information in a time series for improved forecasting performance.

The mathematical formulation of a Bi-LSTM model for time series forecasting is as follows:

Forward LSTM:

$$
\begin{aligned}
i_t^{\rightarrow} &= \sigma(W_{xi}^{\rightarrow} x_t + W_{hi}^{\rightarrow} h_{t-1}^{\rightarrow} + W_{ci}^{\rightarrow} c_{t-1}^{\rightarrow} + b_i^{\rightarrow}) \\
f_t^{\rightarrow} &= \sigma(W_{xf}^{\rightarrow} x_t + W_{hf}^{\rightarrow} h_{t-1}^{\rightarrow} + W_{cf}^{\rightarrow} c_{t-1}^{\rightarrow} + b_f^{\rightarrow}) \\
c_t^{\rightarrow} &= f_t^{\rightarrow} \odot c_{t-1}^{\rightarrow} + i_t^{\rightarrow} \odot \tanh(W_{xc}^{\rightarrow} x_t + W_{hc}^{\rightarrow} h_{t-1}^{\rightarrow} + b_c^{\rightarrow}) \\
o_t^{\rightarrow} &= \sigma(W_{xo}^{\rightarrow} x_t + W_{ho}^{\rightarrow} h_{t-1}^{\rightarrow} + W_{co}^{\rightarrow} c_t^{\rightarrow} + b_o^{\rightarrow}) \\
h_t^{\rightarrow} &= o_t^{\rightarrow} \odot \tanh(c_t^{\rightarrow})
\end{aligned}
$$

Backward LSTM:

$$\begin{aligned}
i_t^{\leftarrow} &= \sigma(W_{xi}^{\leftarrow} x_t + W_{hi}^{\leftarrow} h_{t+1}^{\leftarrow} + W_{ci}^{\leftarrow} c_{t+1}^{\leftarrow} + b_i^{\leftarrow}) \\
f_t^{\leftarrow} &= \sigma(W_{xf}^{\leftarrow} x_t + W_{hf}^{\leftarrow} h_{t+1}^{\leftarrow} + W_{cf}^{\leftarrow} c_{t+1}^{\leftarrow} + b_f^{\leftarrow}) \\
c_t^{\leftarrow} &= f_t^{\leftarrow} \odot c_{t+1}^{\leftarrow} + i_t^{\leftarrow} \odot \tanh(W_{xc}^{\leftarrow} x_t + W_{hc}^{\leftarrow} h_{t+1}^{\leftarrow} + b_c^{\leftarrow}) \\
o_t^{\leftarrow} &= \sigma(W_{xo}^{\leftarrow} x_t + W_{ho}^{\leftarrow} h_{t+1}^{\leftarrow} + W_{co}^{\leftarrow} c_t^{\leftarrow} + b_o^{\leftarrow}) \\
h_t^{\leftarrow} &= o_t^{\leftarrow} \odot \tanh(c_t^{\leftarrow})
\end{aligned}$$

where x_t represents the input at time step t, $h_t^{\rightarrow})$ and $h_t^{\leftarrow}$ represent the hidden states computed by the forward and backward LSTMs respectively, $c_t^{\rightarrow}$ and $c_t^{\leftarrow}$ represent the cell states, and $\sigma(\cdot)$ represents the sigmoid activation function.

To obtain the final hidden state at each time step, the forward and backward hidden states are concatenated:

$$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$$

where [; denotes concatenation.

The final output of the Bi-LSTM model can be calculated as:

$$y_t = g(W_{hy} h_t + b_y)$$

where W_{hy} and b_y represent the weight matrix and bias vector for the output layer, and $g(\cdot)$ represents the activation function.

During training, the parameters of the Bi-LSTM model are optimized by minimizing a suitable loss function using gradient-based optimization algorithms. The gradients are computed through backpropagation through time (BPTT) to update the weights and biases. By considering both past and future information in the time series, Bi-LSTM models can more comprehensively capture temporal dependencies in the time series, thereby improving forecasting performance. The Bi-LSTM model has been widely used in various time series forecasting tasks.

4.3.6 Bidirectional Gated Recurrent Unit(Bi-GRU)

Bi-GRU (Bidirectional Gated Recurrent Unit) is a variant of Recurrent Neural Networks (RNNs) that combines the advantages of GRU and Bidirectional RNNs for time series forecasting tasks. It aims to capture both past and future information in a time series, allowing for a more comprehensive understanding of temporal dependencies.

The mathematical formulation of a Bi-GRU model for time series forecasting is as follows:

Forward GRU:

$$\begin{aligned}
z_t^{\rightarrow} &= \sigma(W_{xz}^{\rightarrow} x_t + W_{hz}^{\rightarrow} h_{t-1}^{\rightarrow} + b_z^{\rightarrow}), \\
r_t^{\rightarrow} &= \sigma(W_{xr}^{\rightarrow} x_t + W_{hr}^{\rightarrow} h_{t-1}^{\rightarrow} + b_r^{\rightarrow}), \\
n_t^{\rightarrow} &= \tanh(W_{xn}^{\rightarrow} x_t + r_t^{\rightarrow} \odot (W_{hn}^{\rightarrow} h_{t-1}^{\rightarrow}) + b_n^{\rightarrow}), \\
h_t^{\rightarrow} &= (1 - z_t^{\rightarrow}) \odot n_t^{\rightarrow} + z_t^{\rightarrow} \odot h_{t-1}^{\rightarrow}.
\end{aligned}$$

Backward GRU:

$$\begin{aligned}
z_t^{\leftarrow} &= \sigma(W_{xz}^{\leftarrow} x_t + W_{hz}^{\leftarrow} h_{t+1}^{\leftarrow} + b_z^{\leftarrow}), \\
r_t^{\leftarrow} &= \sigma(W_{xr}^{\leftarrow} x_t + W_{hr}^{\leftarrow} h_{t+1}^{\leftarrow} + b_r^{\leftarrow}), \\
n_t^{\leftarrow} &= \tanh(W_{xn}^{\leftarrow} x_t + r_t^{\leftarrow} \odot (W_{hn}^{\leftarrow} h_{t+1}^{\leftarrow}) + b_n^{\leftarrow}), \\
h_t^{\leftarrow} &= (1 - z_t^{\leftarrow}) \odot n_t^{\leftarrow} + z_t^{\leftarrow} \odot h_{t+1}^{\leftarrow}.
\end{aligned}$$

where x_t represents the input at time step t, $h_t^{\rightarrow}$ and $h_t^{\leftarrow}$ denote the hidden states computed by the forward and backward GRUs, respectively. $z_t^{\rightarrow}$ and $z_t^{\leftarrow}$ are the update gates, which control the information flow from the previous hidden state to the current hidden state. $r_t^{\rightarrow}$ and $r_t^{\leftarrow}$ are the reset gates, which determine how much of the previous hidden state should be forgotten when computing the candidate hidden state. $n_t^{\rightarrow}$ and $n_t^{\leftarrow}$ represent the candidate hidden states, and $\sigma(\cdot)$ denotes the sigmoid activation function. The updated hidden states are obtained by combining the candidate hidden states with the previous hidden states weighted by the update gates.

To obtain the final hidden state at each time step, the forward and backward hidden states are concatenated:

$$h_t = [h_t^{\rightarrow}; h_t^{\leftarrow}]$$

where [;denotes concatenation.

The final output of the Bi-GRU model can be calculated as:

$$y_t = g(W_{hy}h_t + b_y)$$

where W_{hy} and b_y represent the weight matrix and bias vector for the output layer, and $g(\cdot)$ represents the activation function.

During training, the parameters of the Bi-GRU model are optimized by minimizing a suitable loss function using gradient-based optimization algorithms. The gradients are computed through backpropagation through time (BPTT) to update the weights and biases.

Bi-GRU models leverage both past and future information in the time series, enabling them to capture more comprehensive temporal dependencies and improve forecasting performance compared to traditional RNNs. They have been widely applied in various time series prediction tasks.

4.4 Forecasting Models in Time Series Applications: Simulation Results

4.4.1 Objective function

In deep learning, Mean Squared Error (MSE) and R-squared (R^2) are commonly used evaluation metrics to assess the performance of regression models. These metrics provide valuable insights into the accuracy and goodness-of-fit of the model predictions.

MSE is a measure of how close the predicted values are to the actual values. It calculates the average squared difference between the predicted values $\hat{y}$ and the true values y in the dataset. The formula for MSE is as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Here, n represents the number of samples in the dataset, $\hat{y}_i$ denotes the predicted value for the i-th sample, and y_i represents the corresponding true value.

A lower MSE indicates better performance, as it signifies that the model's predictions are closer to the actual values on average. However, MSE does not provide an intuitive understanding of the proportion of variance explained by the model.

R^2, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that can be explained by the independent variables in a regression model. It quantifies how well the model fits the data compared to a simple baseline model that predicts the mean of the dependent variable. The mathematical formula for R^2 is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

In this equation, y_i represents the true value for the i-th sample, $\hat{y}_i$ denotes the predicted value, and $\bar{y}$ represents the mean of the true values.

R^2 ranges between 0 and 1, where a value of 1 indicates a perfect fit of the model to the data. A higher R^2 suggests that a larger proportion of the variance in the dependent variable can be explained by the independent variables.

Both MSE and R^2 are valuable metrics in deep learning for regression tasks. MSE provides a measure of the average prediction error, while R^2 offers an understanding of the goodness-of-fit of the model.

4.4.2 Data Visualization and Variable Analysis

This study utilizes a dataset comprising 35,064 observations, which has been partitioned into three distinct subsets: the training set, validation set, and test set. The dataset consists of 16 variables, among which two are categorical in nature. The categorical variables encompass "Rain," indicating the occurrence of rainfall, and 'wd' (The data distribution is shown in Figure 4.6), denoting wind direction.

The density distribution plots of the remaining numerical variables are shown in the figure 4.7.

TEMP here represents the temperature (degrees Celsius).

DEWP (Dew Point Temperature) is a variable commonly used in the field of air quality. It refers to the temperature at which air becomes saturated with water vapor at constant atmospheric pressure. When the air reaches its dew point temperature, condensation occurs, leading to the formation of dew or droplets. DEWP is often used as an indicator of humidity in the air.

WSPM (Wind Speed) is another important variable in air quality studies. It

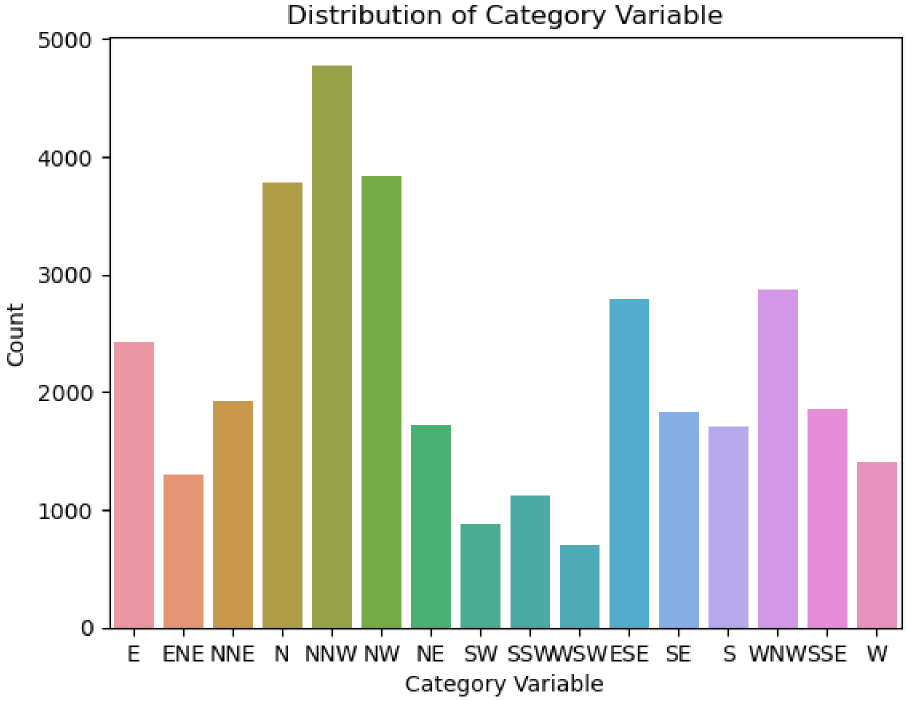


Figure 4.6: Distribution of wd

represents the average speed of wind passing through a fixed point within a given unit of time. Wind speed is typically measured in meters per second (m/s). It serves as a crucial metric for assessing the strength and velocity of wind, and its analysis helps understand the impact of air movement on air quality.

PRES (Pressure) refers to atmospheric pressure, which plays a significant role in air quality research. Atmospheric pressure quantifies the force exerted by the atmosphere on the Earth's surface or any other object per unit area. It arises from the gravitational effects within the Earth's atmosphere and is an essential factor for describing atmospheric conditions.

The analysis reveals several significant correlations among the variables. Figure 4.8 illustrates the relationships observed between PM10 and both CO and NO2, indicating a strong correlation between these variables. Similarly, a notable correlation is observed between CO and NO2. Additionally, the data exhibits a strong correlation between TEMP and O3, as well as between DEMP and TEMP.

The correlation coefficients for these pairs of variables surpass the threshold of 0.6, indicating a substantial linear relationship between them. However, it should be noted that all other variables demonstrate independence from one another, lacking significant correlations.

To address the issue of strong correlation between variables, regularization terms

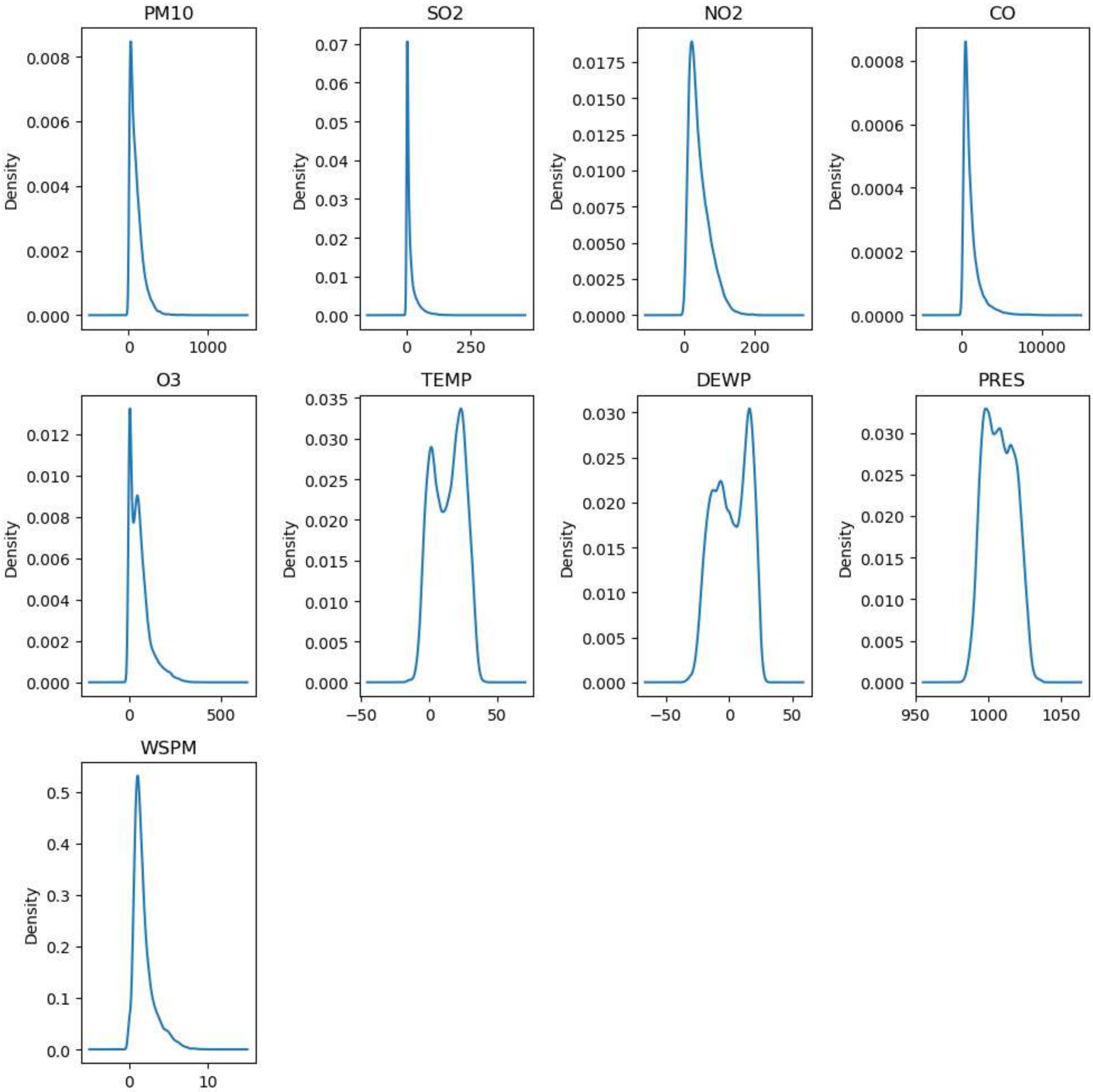


Figure 4.7: Density Plots of Variables

are employed as a means of constraining the magnitudes of model parameters, thereby mitigating the risk of overfitting. This regularization technique aids in preventing excessive reliance on specific features and encourages a more generalizable model.

In addition to regularization, preprocessing techniques were applied to the data prior to model training. This preprocessing step involved normalizing the data, which effectively rescales the values across variables. By normalizing the data, the impact of correlation is reduced, allowing for a more robust analysis and modeling

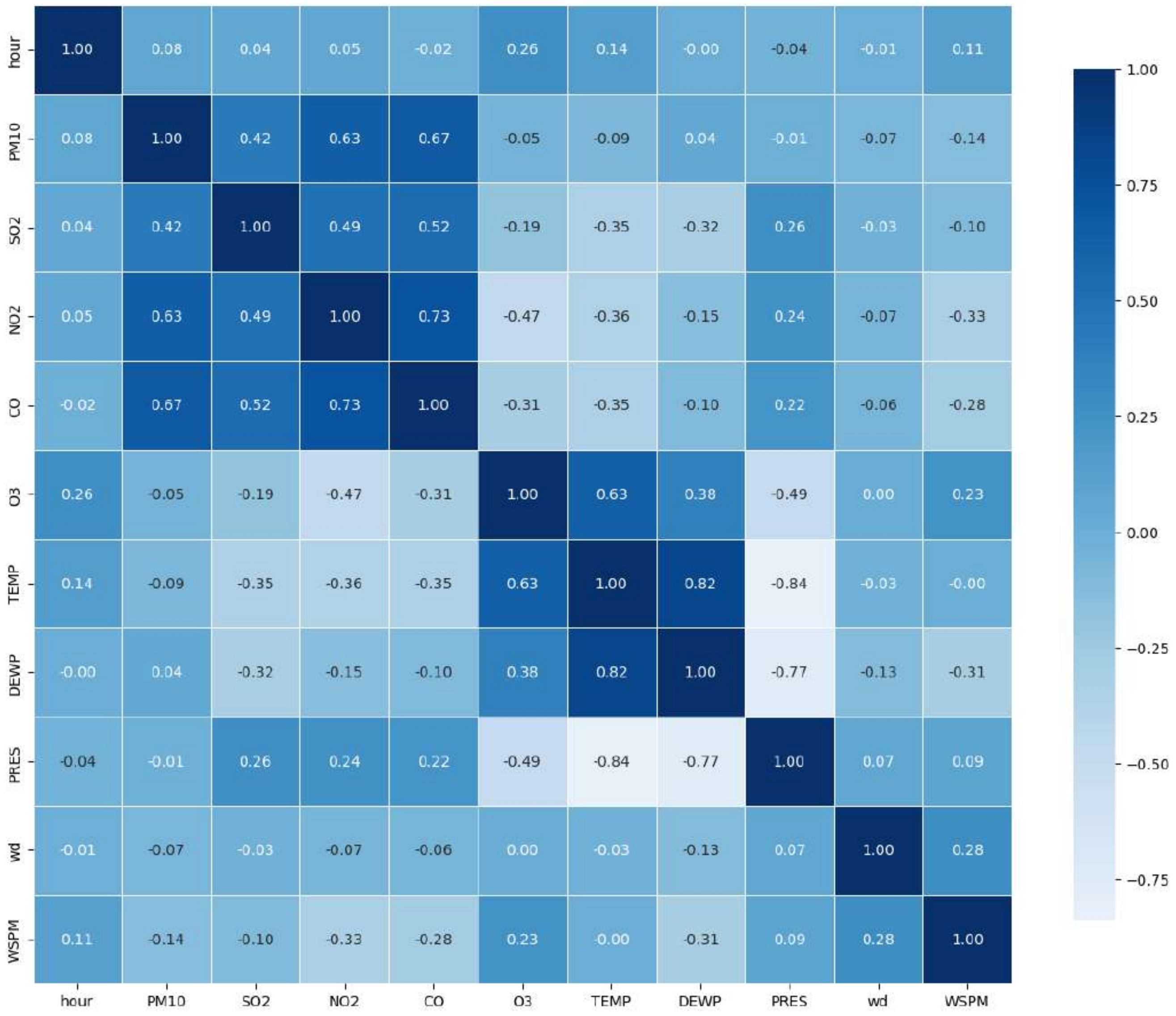


Figure 4.8: Correlation Heatmap of Variables

process.

4.4.3 Simulation Results

Investigating the RNN-Based Model Architecture for Predictive Analysis. Layer Hierarchy(in Table 4.1): By observing the "Layer (type)" column, we can understand the layer hierarchy of the model. In this particular model, there is a SimpleRNN layer with an output shape of (None, 1, 100). This is followed by a Dropout layer, which has the same output shape as the SimpleRNN layer. Then, another SimpleRNN layer appears with an output shape of (None, 50). Finally, there is a Dense layer with an output shape of (None, 1).

Parameter Count: Based on the "Param " column, we can determine the param-

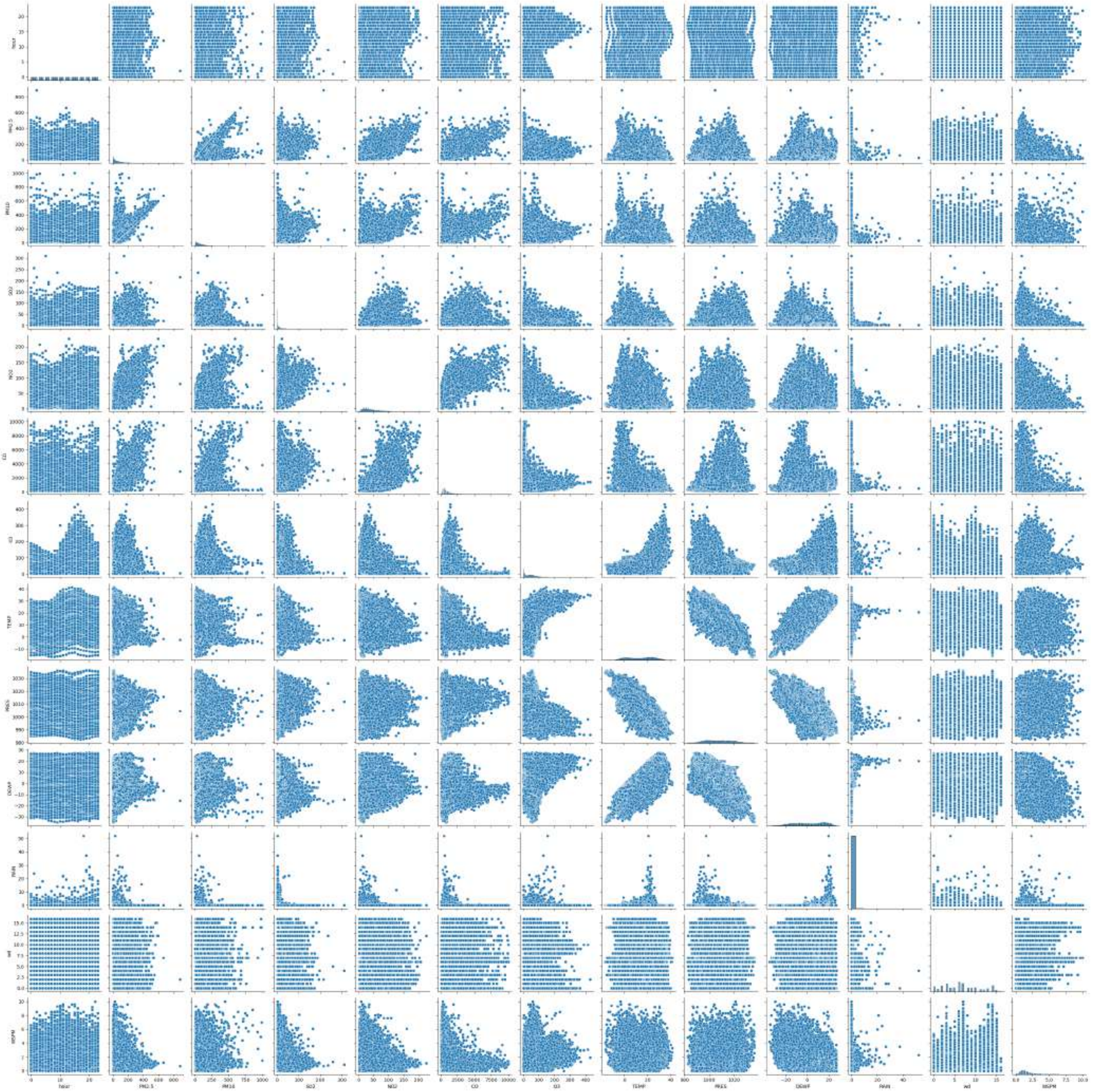


Figure 4.9: Scatter plot of Variables

eter count for each layer. In the table, SimpleRNN layer 1 has 11,700 parameters, while Dropout layer 1 has no trainable parameters. SimpleRNN layer 2 has 7,550 parameters, and Dropout layer 2 also has no trainable parameters. The final Dense layer has 51 parameters. In total, the model has 19,301 parameters, all of which are trainable.

Trainable and Non-trainable Parameters: The "Trainable params" and "Non-trainable params" columns allow us to distinguish between trainable and non-trainable parameters in the model. In this table, all 19,301 parameters are trainable,

Table 4.1: Analysis of Layer Hierarchy and Parameter Count in a SimpleRNN-Based Model

Layer (type)	Output Shape	Param #
simple_rnn_10 (SimpleRNN)	(None, 1, 100)	11,700
dropout_9 (Dropout)	(None, 1, 100)	0
simple_rnn_11 (SimpleRNN)	(None, 50)	7,550
dropout_10 (Dropout)	(None, 50)	0
dense_4 (Dense)	(None, 1)	51
Total params:		**19,301**
Trainable params:		**19,301**
Non-trainable params:		**0**

and there are no non-trainable parameters.

In summary, the model utilizes a simple recurrent neural network (SimpleRNN) architecture and incorporates Dropout layers to reduce the risk of overfitting. The output shape of the model gradually transitions from (None, 1, 100) to (None, 50), and finally becomes (None, 1). There are a total of 19,301 trainable parameters in the model, which play a crucial role in capturing the relationships between the input data and the target variable during training.

Table 4.2: Exploring LSTM-Based Model Architecture for Predictive Analysis

Layer (type)	Output Shape	Param #
lstm (LSTM)	(None, 1, 100)	46,800
dropout_11 (Dropout)	(None, 1, 100)	0
lstm_1 (LSTM)	(None, 50)	30,200
dropout_12 (Dropout)	(None, 50)	0
dense_5 (Dense)	(None, 1)	51
Total params:		**77,051**
Trainable params:		**77,051**
Non-trainable params:		**0**

Investigating the LSTM-Based Model Architecture for Predictive Analysis The table 4.2 presents information about a model architecture based on Long Short-Term Memory (LSTM) layers. The model consists of two LSTM layers, a Dropout layer, and a Dense layer.

LSTM Layer 1: This layer has an output shape of (None, 1, 100), indicating that it produces a sequence of vectors with a length of 1 and each vector having 100 dimensions. It contains 46,800 parameters.

Dropout Layer: The Dropout layer helps reduce overfitting by randomly setting a fraction of the input units to 0 during training. It does not alter the output shape and has no trainable parameters.

LSTM Layer 2: This layer has an output shape of (None, 50), representing a sequence of vectors with a length of 50. It contains 30,200 parameters.

Dense Layer: The Dense layer performs a linear transformation on the input data and has an output shape of (None, 1), indicating a single-dimensional output. It contains 51 parameters.

Overall, the model has a total of 77,051 parameters, all of which are trainable. These parameters capture the relationships between the input data and the target variable during the training process.

Table 4.3: GRU-Based Model Architecture for Predictive Analysis

Layer (type)	**Output Shape**	**Param #**
gru_1 (GRU)	(None, 1, 100)	35,400
dropout_14 (Dropout)	(None, 1, 100)	0
gru_2 (GRU)	(None, 50)	22,800
dropout_15 (Dropout)	(None, 50)	0
dense_6 (Dense)	(None, 1)	51
Total params:		**58,251**
Trainable params:		**58,251**
Non-trainable params:		**0**

Investigating the GRU-Based Model Architecture for Predictive Analysis. The Table 4.3 presents information about a model architecture based on Gated Recurrent Unit (GRU) layers. The model consists of two GRU layers, a Dropout layer, and a Dense layer.

GRU Layer 1: This layer has an output shape of (None, 1, 100), indicating that it produces a sequence of vectors with a length of 1 and each vector having 100 dimensions. It contains 35,400 parameters.

Dropout Layer: The Dropout layer helps reduce overfitting by randomly setting a fraction of the input units to 0 during training. It does not alter the output shape and has no trainable parameters.

GRU Layer 2: This layer has an output shape of (None, 50), representing a sequence of vectors with a length of 50. It contains 22,800 parameters.

Dense Layer: The Dense layer performs a linear transformation on the input data and has an output shape of (None, 1), indicating a single-dimensional output. It contains 51 parameters.

Overall, the model has a total of 58,251 parameters, all of which are trainable. These parameters capture the relationships between the input data and the target variable during the training process.

Table 4.4: Bidirectional RNN Model Architecture for Sequence Prediction

Layer (type)	**Output Shape**	**Param #**
bidirectional_1 (Bidirectional)	(None, 1, 200)	23,400
dropout_1 (Dropout)	(None, 1, 200)	0
dense_1 (Dense)	(None, 1, 1)	201
Total params:		**23,601**
Trainable params:		**23,601**
Non-trainable params:		**0**

Investigating the Bi-RNN-Based Model Architecture for Predictive Analysis The Table 4.4 presents information about a bidirectional RNN model architecture. The model consists of a Bidirectional layer, a Dropout layer, and a Dense layer.

Bidirectional Layer: This layer uses two separate RNN layers to process the input sequence in both forward and backward directions. It has an output shape of (None, 1, 200), indicating that it produces a sequence of vectors with a length of 1 and each vector having 200 dimensions. It contains 23,400 parameters.

Dropout Layer: The Dropout layer helps reduce overfitting by randomly setting a fraction of the input units to 0 during training. It does not alter the output shape and has no trainable parameters.

Dense Layer: The Dense layer performs a linear transformation on the input data and has an output shape of (None, 1, 1), indicating a single-dimensional output. It contains 201 parameters.

Overall, the model has a total of 23,601 parameters, all of which are trainable. These parameters capture the relationships between the input sequence and the target variable during the training process.

Investigating the BiLSTM-Based Model Architecture for Predictive Analysis The Table 4.5 presents information about a bidirectional Long Short-Term Memory (LSTM) model architecture. The model consists of a Bidirectional layer, a Dropout layer, and a Dense layer.

Bidirectional Layer: This layer utilizes two separate LSTM layers to process the input sequence in both forward and backward directions. It has an output shape of

Table 4.5: Bidirectional LSTM Model Architecture

Layer (type)	**Output Shape**	**Param #**
bidirectional_2 (Bidirectional)	(None, 1, 200)	93,600
dropout_2 (Dropout)	(None, 1, 200)	0
dense_2 (Dense)	(None, 1, 1)	201
Total params:		**93,801**
Trainable params:		**93,801**
Non-trainable params:		**0**

(None, 1, 200), indicating that it produces a sequence of vectors with a length of 1 and each vector having 200 dimensions. It contains 93,600 parameters.

Dropout Layer: The Dropout layer helps reduce overfitting by randomly setting a fraction of the input units to 0 during training. It does not alter the output shape and has no trainable parameters.

Dense Layer: The Dense layer performs a linear transformation on the input data and has an output shape of (None, 1, 1), indicating a single-dimensional output. It contains 201 parameters.

Overall, the model has a total of 93,801 parameters, all of which are trainable. These parameters capture the relationships between the input sequence and the target variable during the training process.

Table 4.6: Bidirectional GRU Model Architecture

Layer (type)	**Output Shape**	**Param #**
bidirectional_3 (Bidirectional)	(None, 1, 200)	70,800
dropout_3 (Dropout)	(None, 1, 200)	0
dense_3 (Dense)	(None, 1, 1)	201
Total params:		**71,001**
Trainable params:		**71,001**
Non-trainable params:		**0**

Investigating the BiGRU-Based Model Architecture for Predictive Analysis The Table 4.6 presents information about a bidirectional GRU model architecture. The model consists of a Bidirectional layer, a Dropout layer, and a Dense layer.

Bidirectional Layer: This layer employs two separate GRU layers to process the input sequence in both forward and backward directions. It has an output shape of (None, 1, 200), indicating that it produces a sequence of vectors with a length of 1 and each vector having 200 dimensions. It contains 70,800 parameters.

Dropout Layer: The Dropout layer is used to reduce overfitting by randomly setting a fraction of the input units to 0 during training. It does not change the output shape and has no trainable parameters.

Dense Layer: The Dense layer performs a linear transformation on the input data and has an output shape of (None, 1, 1), representing a single-dimensional output. It contains 201 parameters.

Overall, the model has a total of 71,001 parameters, all of which are trainable. These parameters capture the relationships between the input sequence and the target variable during the training process.

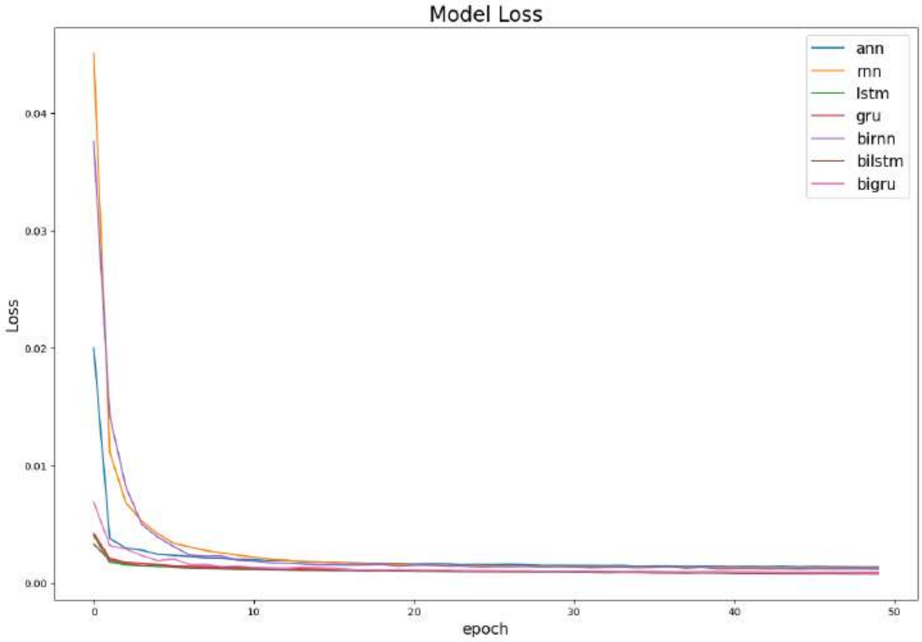


Figure 4.10: Learning curves of forecast models.

The figure 4.10 depicts the learning curves of various models including ANN, RNN, LSTM, GRU, BiRNN, BiLSTM, and BiGRU. The curves exhibit a discernible decline, suggesting a notable improvement in performance as the number of training samples processed increases. This observed trend serves as an encouraging indication that the models are effectively learning from the provided data and successfully adapting to the given task.

The figure 4.10 demonstrates that all considered models exhibit a consistent pattern of enhanced performance with increasing amounts of training data. This positive trend underscores the models' capability to effectively learn and adapt.

To ensure a thorough and comprehensive comparison of various models, it is imperative to engage in a meticulous analysis of quantitative indicators. By carefully

Table 4.7: Forecast Quality of Current PM2.5.

	R^2	**MSE**	**TimeSpent**
ANN	0.9085	0.0013	**6.91s**
RNN	0.9177	0.0013	21.9s
LSTM	0.8970	0.0018	38.2s
GRU	0.9080	0.0015	38.2s
Bi-RNN	**0.9212**	**0.0012**	20s
Bi-LSTM	0.9086	0.0014	36.9s
Bi-GRU	0.9136	0.0015	32.8 s

examining these metrics, one can effectively assess the strengths and limitations of each model relative to a specific problem at hand. Such a detailed evaluation enables researchers and practitioners to make informed decisions regarding the most suitable model for their intended application.

Moreover, considering the available computational resources plays a crucial role in this comparative analysis. It is essential to evaluate the performance of each model while taking into account the trade-off between computational time and predictive power. This assessment helps in determining the optimal deployment model that strikes a balance between performance and efficiency. Finding the right equilibrium between computational requirements and the model's ability to generate accurate predictions is vital for real-world applications where computational resources are often limited.

By conducting a comprehensive analysis of quantitative indicators and considering the computational trade-offs, researchers and practitioners can make informed choices when selecting a model for a specific task. This approach ensures that the chosen model aligns with both the desired level of performance and the available computational resources, ultimately enhancing the effectiveness and efficiency of the overall system.

Model Performance Comparison: By observing the R^2 and MSE values, we can compare the predictive performance of different machine learning models. Based on the table data, the Bi-RNN model has the highest R^2 value (0.9212), indicating the best fitting performance in predicting the current PM2.5 quality. Additionally, the ANN model and Bi-GRU model also exhibit relatively high R^2 values (0.9085 and 0.9136).

Prediction Error Comparison: The MSE values allow us to evaluate the average squared error of different models. A smaller MSE value indicates less difference between the predicted values and the true values. From the table, it can be observed that the Bi-RNN model and the ANN model have the lowest MSE values (0.0012 and 0.0013), implying smaller errors in their predictions.

Training Time Comparison: The TimeSpent column in the table displays the training time for each model. We observe that the ANN model requires the least amount of time for training (6.91 seconds), while the LSTM model and Bi-LSTM model have relatively longer training times (38.2 seconds and 36.9 seconds respectively). Thus, in practical applications, there is a trade-off between model performance and training time.

Based on the above analysis, we can draw some preliminary conclusions: The Bi-RNN model performs exceptionally well in the current PM2.5 prediction task, showing higher predictive accuracy and lower error. Although the ANN model has the shortest training time, its predictive performance slightly lags behind other models. Additionally, the LSTM and GRU models also demonstrate good performance, although their predictive accuracy is slightly lower compared to the Bi-RNN model.

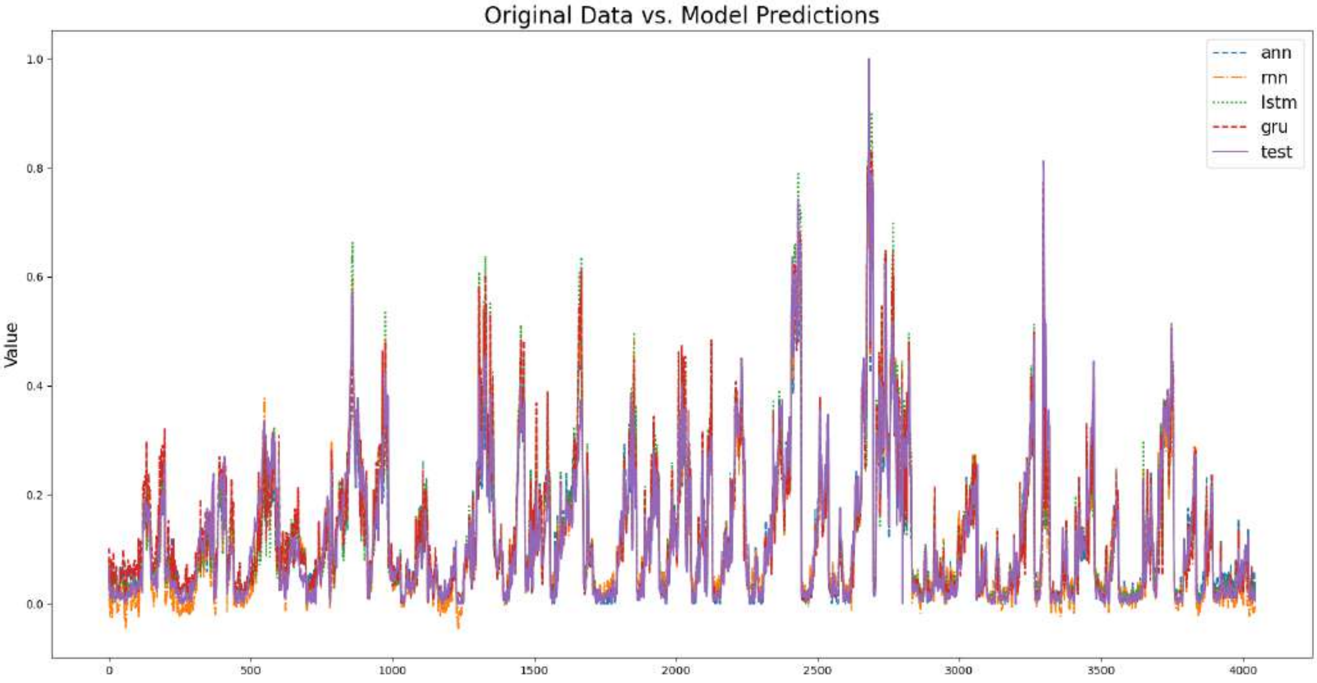


Figure 4.11: Comparison of Raw and Predicted Data in Time Series Forecasting

Figure 4.11 and Figure 4.12 present line charts illustrating the performance of seven different models. In Figure 4.11, five lines depict the actual test values alongside the predicted values generated by the ANN, RNN, LSTM, and GRU models.

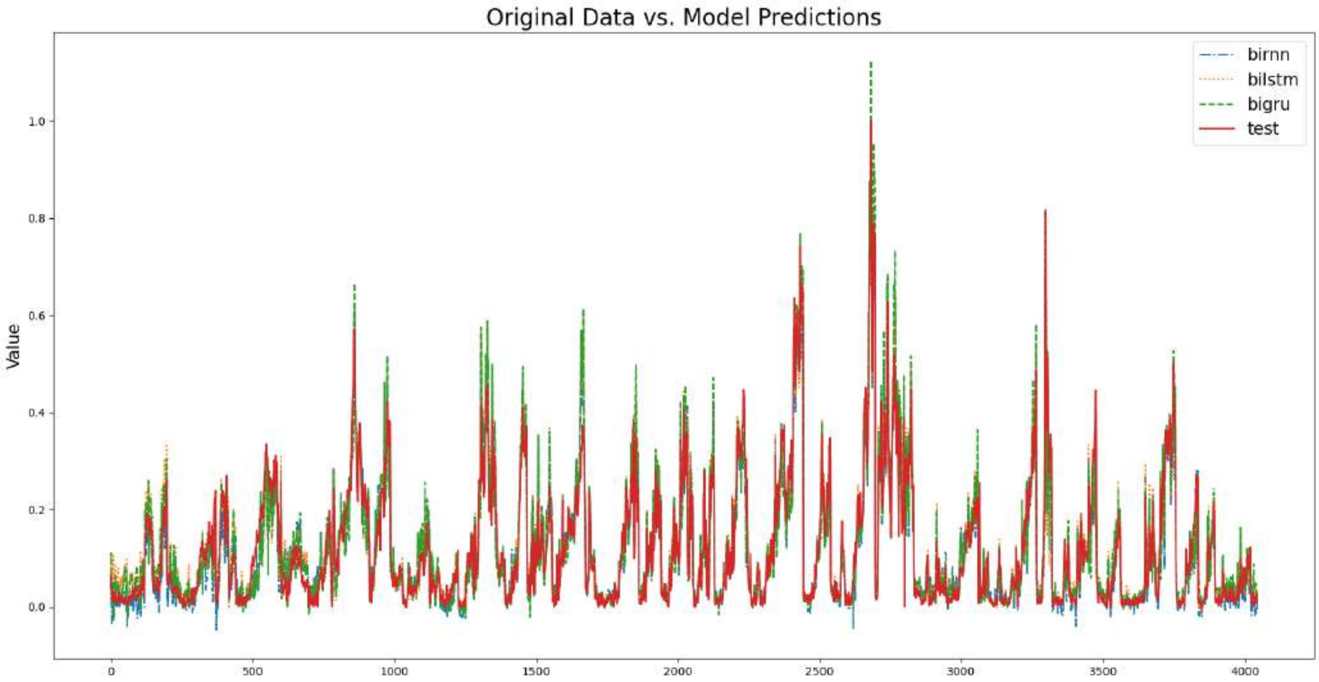


Figure 4.12: Comparison of Raw and Predicted Data in Time Series Forecasting

Conversely, Figure 4.12 comprises four line charts representing the actual test values together with the predicted values from the BiRNN, BiLSTM, and BiGRU models.

The visual examination of these figures indicates that all seven models have achieved favorable outcomes. The disparities between the predicted results and the original values are minimal, suggesting a high level of accuracy in the models' predictions. Furthermore, the trend changes observed in the predictions closely align with the patterns exhibited by the original data.

This consistency between the predicted and actual values signifies the models' ability to effectively capture and replicate the underlying trends and patterns within the dataset. It implies that these models possess a robust learning capability and can generate reliable predictions that closely resemble the true values.

4.5 Conclusion to Chapter 4

In this chapter, we delved into the standardized usage procedures of machine learning for time series prediction of PM2.5, an important environmental factor with significant implications for air quality assessment and public health. Through a rigorous analysis and comparison of seven different models, namely ANN, RNN, LSTM, GRU, Bi-RNN, Bi-LSTM, and Bi-GRU, we gained valuable insights into their performance and predictive capabilities.

The foundation of our methodology lies in the division of the dataset into three distinct subsets: the training set, verification set, and test set. This approach ensures the integrity and reliability of our model evaluation by providing separate data for training, validation, and final testing. By adhering to these standardized procedures, we were able to derive accurate and meaningful conclusions about the performance of each model.

Our evaluation employed two key metrics, namely R^2 (coefficient of determination) and MSE (mean squared error), to comprehensively compare the predictive abilities of the different models. The R^2 value serves as a measure of how well a model fits the observed data, indicating its ability to capture the underlying patterns and trends within the PM2.5 time series. On the other hand, MSE provides a quantitative assessment of the average squared difference between the predicted values and the true values, offering insights into the overall accuracy of the models' predictions.

After comparing the prediction results with the actual PM2.5 data, we found that all seven models exhibited excellent simulation outcomes. However, a closer examination revealed nuanced differences in their performance. The standout performer was the Bi-RNN model, which demonstrated the highest R^2 value of 0.9212. This exceptional result signifies its strong fitting performance and suggests that it is particularly adept at capturing the intricate dynamics of PM2.5 time series. Furthermore, both the ANN and Bi-GRU models also displayed commendable predictive capabilities, with relatively high R^2 values of 0.9085 and 0.9136 respectively.

To assess the models' prediction errors, we examined their MSE values. The Bi-RNN model and the ANN model stood out as frontrunners in this regard, showcasing the lowest MSE values of 0.0012 and 0.0013 respectively. These smaller errors indicate a closer alignment between the predicted values and the true values, further affirming the accuracy and reliability of these models' predictions.

Additionally, we took into consideration the training time required for each model. The ANN model exhibited the shortest training time at 6.91 seconds, highlighting its computational efficiency. Conversely, the LSTM and Bi-LSTM models necessitated longer training times of 38.2 seconds and 36.9 seconds respectively. This observation underscores the trade-off that often exists between model performance and computational resources.

In conclusion, our comprehensive analysis offers valuable insights into the standardized usage procedures of machine learning for PM2.5 prediction. The findings highlight the superior performance of the Bi-RNN model, particularly in terms of fitting accuracy and lower prediction errors. However, the ANN, LSTM, GRU, Bi-LSTM, and Bi-GRU models also demonstrated competitive predictive capabilities, albeit with slightly varying degrees of accuracy and computational requirements.

These research findings contribute to the field of environmental data analysis by providing guidance on the selection and deployment of appropriate machine learning models for PM2.5 prediction tasks. Researchers and practitioners can utilize these insights to make informed decisions based on their specific requirements, striking a balance between prediction accuracy, computational efficiency, and training time. Ultimately, our study advances the understanding and application of machine learning techniques in environmental monitoring and enhances our ability to assess and mitigate air pollution-related risks.

Chapter 5

Ensemble Learning Methods for air quality evaluation system

In this section, we delve into the application of ensemble learning methods for forecasting air quality time series. Specifically, we examine three widely used ensemble learning models, namely XGBoost, LightGBM, and CatBoost.

To ensure consistency and comparability with the previous chapter, we employ the same dataset and adopt similar data processing techniques. This entails steps such as outlier removal, missing value imputation, and data smoothing, which contribute to enhancing the model's accuracy and stability.

By leveraging these three ensemble learning models, we can approach air quality time series modeling and prediction from diverse perspectives. Each model possesses distinct advantages and is suitable for specific scenarios. In practical applications, the selection of an appropriate model depends on the particular context. The utilization of ensemble learning allows for the amalgamation of predictions from multiple weaker learners, thereby improving overall predictive performance and exhibiting robustness.

Some algorithms discussed in this chapter have been utilized by the author in their published works [102] and [103].

5.1 Methodology

5.1.1 XGBoost (Extreme Gradient Boosting)

XGBoost (Extreme Gradient Boosting) is an ensemble learning model that has gained popularity for its effectiveness in time series forecasting tasks. It combines the principles of gradient boosting and decision trees to achieve accurate predictions on temporal data.

The XGBoost algorithm can be formulated as follows:

Given a training dataset $(\mathbf{x}_i, y_i)i = 1^N$, where $\mathbf{x}_i$ represents the feature vector at time step i and y_i denotes the corresponding target value, the objective of XGBoost is to find a prediction function $F(\mathbf{x})$ that minimizes the regularized objective function defined as:

$$Obj = \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)) + \sum k = 1^K \Omega(f_k)$$

where L is the loss function measuring the discrepancy between the predicted values and the actual targets, $F(\mathbf{x}_i)$, and $\Omega(f_k)$ is the regularization term penalizing complex models. Here, f_k represents individual decision trees in the ensemble.

To iteratively build the ensemble, XGBoost employs a boosting strategy, which involves sequentially adding new weak learners to improve upon the residuals left by previous models. The prediction function at each iteration is given by the sum of all the individual tree predictions:

$$F_t(\mathbf{x}) = \sum_{k=1}^{t} f_k(\mathbf{x})$$

where t denotes the current iteration.

The key idea behind XGBoost lies in the optimization of the objective function through gradient descent. By computing the gradients of the loss function with respect to the ensemble's predictions and using them to fit a new decision tree, XGBoost learns how to correct the mistakes made by previous models. This process is repeated iteratively until the objective function is minimized.

To prevent overfitting and enhance model generalization, XGBoost incorporates regularization into the objective function. The regularization term $\Omega(f_k)$ controls the complexity of individual trees by penalizing their structure or leaf weights. This prevents the model from becoming too complex and improves its ability to generalize to unseen data.

XGBoost also includes additional advanced features such as handling missing values, subsampling, and column sampling. These techniques further enhance the model's performance and robustness.

In summary, XGBoost is a powerful ensemble learning algorithm that combines gradient boosting and decision trees for time series forecasting. It optimizes a regularized objective function through iterative training, leveraging the strengths of weak learners to make accurate predictions.

5.1.2 LightGBM

LightGBM is a gradient boosting framework that has gained popularity for its efficient and effective performance in time series forecasting tasks. It is specifically designed to handle large-scale datasets and provides superior accuracy while maintaining fast training times.

The LightGBM algorithm can be described as follows:

Given a training dataset $(\mathbf{x}_i, y_i)i = 1^N$, where $\mathbf{x}_i$ represents the feature vector at time step i and y_i denotes the corresponding target value, the objective of LightGBM is to find a prediction function $F(\mathbf{x})$ that minimizes the following loss function:

$$Obj = \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)) + \sum k = 1^K \Omega(f_k)$$

Here, L is the loss function that measures the discrepancy between the predicted values and the actual targets, $F(\mathbf{x}_i)$. The regularization term $\Omega(f_k)$ penalizes complex models to prevent overfitting. Similar to XGBoost, f_k represents individual decision trees in the ensemble.

LightGBM uses a leaf-wise tree growth strategy, which differs from traditional level-wise approaches. In leaf-wise growth, the algorithm grows the tree by splitting the leaf that will result in the largest reduction in the loss function. This strategy leads to faster convergence and better overall performance.

To handle time series data, LightGBM includes a special feature called "categorical feature support". It can effectively handle categorical features without the need for one-hot encoding, reducing memory usage and computational complexity.

Furthermore, LightGBM incorporates additional techniques such as feature subsampling and bagging, which improve the model's generalization ability and robustness. Feature sub-sampling randomly selects a subset of features for each tree, reducing overfitting and enhancing model diversity. Bagging involves training multiple models on different subsets of the training data and averaging their predictions, further improving prediction accuracy.

LightGBM also employs histogram-based algorithms to speed up the training process by grouping values into discrete bins. This technique reduces the memory usage and allows for faster computation.

LightGBM is a powerful gradient boosting framework for time series forecasting. It minimizes a loss function using a leaf-wise tree growth strategy and incorporates regularization techniques to prevent overfitting. With its efficient handling of large-scale datasets and support for categorical features, LightGBM provides accurate predictions with fast training times.

5.1.3 CatBoost

CatBoost is a gradient boosting algorithm that has gained popularity for its ability to handle categorical variables effectively in time series forecasting tasks. It incorporates specific techniques to handle categorical features and provides robust predictions.

The CatBoost algorithm can be described as follows:

Given a training dataset $(\mathbf{x}_i, y_i)i = 1^N$, where $\mathbf{x}_i$ represents the feature vector at time step i and $y_i$denotes the corresponding target value, the objective of CatBoost is to find a prediction function $F(\mathbf{x})$ that minimizes the following loss function:

$$Obj = \sum_{i=1}^{N} L(y_i, F(\mathbf{x}_i)) + \sum k = 1^K \Omega(f_k)$$

Here, L is the loss function that measures the discrepancy between the predicted values and the actual targets, $F(\mathbf{x}_i)$. The regularization term $\Omega(f_k)$ penalizes complex models to prevent overfitting. Similar to XGBoost and LightGBM, f_k represents individual decision trees in the ensemble.

CatBoost introduces an innovative technique called "Ordered Boosting" to deal with categorical features directly. It builds a separate decision tree for each categorical feature using an ordered method that takes into account the statistical properties of the categories. This approach enables CatBoost to capture valuable information from categorical features and make accurate predictions.

To improve generalization and avoid overfitting, CatBoost employs a combination of gradient-based optimization and random permutations. It randomly permutes the order of the categorical values during training to reduce the impact of the order bias.

CatBoost also incorporates a novel method called "Taylor Series Expansion" to approximate the target function. This technique helps to model the nonlinear relationships between input features and the target variable more accurately.

In addition, CatBoost includes techniques such as learning rate scheduling, feature sub-sampling, and early stopping. Learning rate scheduling adjusts the learning rate during training to improve convergence and avoid overshooting. Feature sub-sampling randomly selects a subset of features for each tree, reducing overfitting and enhancing model diversity. Early stopping stops the training process when the model's performance on a validation set no longer improves, preventing overfitting and saving computational resources.

CatBoost is a powerful gradient boosting algorithm designed for time series forecasting tasks with categorical features. It incorporates ordered boosting, random permutations, and Taylor Series Expansion to handle categorical variables effectively. With its regularization techniques and additional features like learning rate scheduling and feature sub-sampling, CatBoost provides robust predictions while avoiding overfitting.

5.1.4 Simulation Results

Table 5.1: XGBoost Model Parameters

Parameter	Value
objective	reg:squarederror
eval_metric	mae
learning_rate	0.3
max_depth	6
subsample	1
random_state	None

Analysis of XGBoost Model Parameters. The Table 5.1 presents the key model parameters for an XGBoost model used in regression tasks. This analysis provides insights into the parameter choices made during model training, enabling researchers to better understand the model's behavior and make informed decisions when applying XGBoost to their own regression problems.

- Objective: The objective function used for regression tasks. The reg: squared-error objective minimizes the mean squared error (MSE) between predicted and actual target values.

- Evaluation Metric: The metric used to evaluate the model's performance during training. The 'mae' metric measures the mean absolute error (MAE), providing insights into the average magnitude of prediction errors.

- Learning Rate: Controls the step size at each boosting iteration. A learning rate of 0.3 indicates relatively large steps, which can expedite convergence but requires careful tuning to avoid overshooting optimal solutions.

- Maximum Depth: Specifies the maximum depth of each decision tree. With a maximum depth of 6, the model can capture complex interactions between features but risks overfitting if not properly regularized.

- Subsample: Controls the fraction of training instances used for each tree. A subsample value of 1 implies using all training instances, potentially resulting in higher variance models. Reducing this value can reduce overfitting.

- Random State: The seed value for random number generation. By setting it to None, the model's behavior will vary across different runs, which is useful for assessing model stability and generalization performance.

Table 5.2: LightGBM Model Parameters

Parameter	Value	Parameter	Value
boosting_type	gbdt	min_child_samples	20
class_weight	None	min_child_weight	0.001
colsample_bytree	1.0	min_split_gain	0.0
importance_type	split	n_estimators	100
learning_rate	0.1	n_jobs	-1
max_depth	-1	num_leaves	31
objective	None	random_state	None
reg_alpha	0.0	reg_lambda	0.0
silent	warn	subsample	1.0
subsample_for_bin	200000	subsample_freq	0

Analysis of LightGBM Model Parameters. The Table 5.2 presents a comprehensive overview of the model parameters used in LightGBM, a popular gradient boosting framework.

Parameter Descriptions:

- Boosting Type: Specifies the type of boosting algorithm utilized. The 'gbdt' (gradient boosting decision tree) method is employed as the default boosting type.
- Learning Rate: Controls the step size at each iteration during boosting. A learning rate of 0.1 indicates moderate steps, balancing convergence speed and accuracy.
- Maximum Depth: Sets the maximum depth of each decision tree. With a value of -1, there is no constraint on the maximum depth, allowing trees to grow without limitation.
- Number of Leaves: Determines the maximum number of leaves in a tree. The default value of 31 ensures sufficient flexibility for capturing complex relationships within the data.
- Regularization: Regulates overfitting through regularization techniques. Both L1 (reg alpha) and L2 (reg lambda) regularization terms are set to zero by default, indicating no regularization.
- Subsampling: Controls the fraction of samples used for each boosting iteration. A subsample value of 1.0 implies using the entire training set, ensuring optimal model performance.

Investigating the CatBoost-Based Model Architecture for Predictive Analysis. The CatBoostRegressor algorithm offers several parameters that can be adjusted to improve the performance of the regression model. Understanding these parameters and their effects is crucial for researchers and practitioners aiming to achieve optimal results. This analysis provides a comprehensive examination of the key parameters utilized in the CatBoostRegressor model.

The CatBoostRegressor model was trained using the following parameter settings:

- nan mode: The "Min" value was chosen, indicating that missing values are treated as minimal values during training.

Table 5.3: CatBoostRegressor Parameters

Parameter	Value	Parameter	Value
nan_mode	Min	feature_border_type	GreedyLogSum
eval_metric	RMSE	bayesian_matrix_reg	0.10000000149011612
iterations	1000	force_unit_auto_pair_weights	False
sampling_frequency	PerTree	l2_leaf_reg	3
leaf_estimation_method	Newton	random_strength	1
grow_policy	SymmetricTree	rsm	1
penalties_coefficient	1	boost_from_average	True
boosting_type	Plain	model_size_reg	0.5
model_shrink_mode	Constant	pool_metainfo_options	{'tags': {}}
depth	6	subsample	0.800000011920929
posterior_sampling	False	use_best_model	False
border_count	254	random_seed	0
auto_class_weights	None	loss_function	RMSE
sparse_features_conflict_fraction	0	learning_rate	0.06794899702072144
leaf_estimation_backtracking	AnyImprovement	score_function	Cosine
best_model_min_trees	1	leaf_estimation_iterations	1
model_shrink_rate	0	bootstrap_type	MVS
min_data_in_leaf	1	max_leaves	64

- eval metric: The evaluation metric selected for the model is RMSE (Root Mean Square Error), which measures the accuracy of the regression predictions.

- iterations: The model was iterated 1000 times to refine the results.

- sampling frequency: PerTree sampling frequency was utilized to determine the random subsampling strategy.

- leaf estimation method: The Newton method was employed to estimate leaf values during the tree construction process.

- grow policy: The SymmetricTree grow policy allows symmetric tree growth from the root.

- penalties coefficient: The penalty coefficient was set to 1, influencing the regularization applied to the model.

- boosting type: The Plain boosting type was chosen, indicating no additional modifications to the standard gradient boosting method.

- model shrink mode: The Constant shrink mode was used, resulting in a constant shrinkage rate during model training.

- feature border type: The GreedyLogSum feature border type was utilized, which enables efficient handling of categorical features.

- bayesian matrix reg: A regularization parameter of 0.1 was applied to the Bayesian matrix.

- force unit auto pair weights: This option was disabled (False), allowing the model to calculate auto-pair weights as needed.

- l2 leaf reg: The L2 regularization coefficient was set to 3, controlling the strength of L2 regularization in the model.

- random strength: Random strength was set to 1, introducing random perturbations to feature values during training.

- rsm: RSM (Random Subspace Method) was set to 1, enabling random subspace selection for each tree.

- boost from average: Boosting from average predictions was enabled, contributing to more stable model training.

- model size reg: A model size regularization of 0.5 was applied, influencing the complexity of the resulting model.

- pool metainfo options: Additional meta-information options were not utilized, and the tags dictionary remained empty.

- subsample: Each tree was trained on a random subsample containing 80% of the training data.

- use best model: The best model was not used during training.

- random seed: A random seed value of 0 was set to ensure reproducibility of results.

- depth: Tree depth was set to 6, limiting the complexity of individual trees and preventing overfitting.

- posterior sampling: Posterior sampling was disabled, indicating that only one sample was considered during tree construction.

- border count: The border count was set to 254, determining the number of buckets for numerical features.

- auto class weights: No automatic class weights were applied.
- sparse features conflict fraction: Sparse features conflict fraction was set to 0, which handles conflicts between sparse features.
- leaf estimation backtracking: The AnyImprovement backtracking method was utilized during leaf value estimation.
- best model min trees: A minimum of 1 tree is required for the best model selection.
- model shrink rate: Model shrinkage rate was set to 0, indicating no shrinkage during training.
- min data in leaf: Each leaf must contain at least 1 data point.
- loss function: The loss function used to optimize the model is RMSE, aligning with the evaluation metric.
- learning rate: A learning rate of 0.06794899702072144 was applied, controlling the step size during optimization.
- score function: Cosine similarity was selected as the score function.
- leaf estimation iterations: Leaf estimation iterations were set to 1, ensuring efficient leaf values estimation.
- bootstrap type: Multiple times with replacement (MVS) bootstrap type was used to construct the trees.
- max leaves: Each tree was allowed a maximum of 64 leaves.

Table 5.4: Forecast Quality of Current PM2.5.

	R^2	**MSE**	**TimeSpent**
XGBoost	0.8551	0.0018	**756 ms**
LightGBM	0.9134	0.0013	140 ms
CatBoosst	0.9112	0.0014	2.96 s

Simulation Results. The table provided presents a comparison of three additional machine learning models, namely XGBoost, LightGBM, and CatBoost, in terms of their forecast quality for current PM2.5 values. The metrics evaluated include R^2 (coefficient of determination), MSE (mean squared error), and TimeSpent (training time).

Starting with the analysis of R^2 values, which indicate the fitting performance of the models, we observe that LightGBM achieved the highest value of 0.9134. This suggests that LightGBM is able to capture a significant portion of the variance in the PM2.5 data, making it a strong contender for accurate predictions. XGBoost and CatBoost also performed well in this regard, with R^2 values of 0.8551 and 0.9112 respectively.

Moving on to the evaluation of prediction errors using MSE, a metric that quantifies the average squared difference between predicted and true values, we find that LightGBM achieved the lowest MSE value of 0.0013. This indicates that LightGBM's predictions have the least deviation from the true PM2.5 values on average. XGBoost and CatBoost also demonstrated relatively low MSE values of 0.0018 and 0.0014 respectively.

Considering training time, as represented by the TimeSpent column, we can observe that LightGBM had the fastest training time at 140 ms. XGBoost required 756 ms, while CatBoost took 2.96 seconds to complete the training process. This implies that LightGBM is computationally efficient, making it a favorable choice for time-sensitive applications.

In summary, based on the analysis of the presented table, LightGBM emerges as the top-performing model among the three, exhibiting the highest R^2 value and the lowest MSE value. Additionally, LightGBM demonstrated the shortest training time, further highlighting its computational efficiency. XGBoost and CatBoost also showcased competitive forecast quality, although with slightly lower R^2 values and MSE values compared to LightGBM.

These findings contribute to the understanding of alternative machine learning models for PM2.5 prediction. Researchers and practitioners can consider employing LightGBM, XGBoost, or CatBoost based on their specific requirements in terms of prediction accuracy, computational efficiency, and time constraints.

Figure 5.1 depicts line charts illustrating the performance of three distinct models, namely XGBoost, LightGBM, and CatBoost. The figure consists of four line charts, representing the actual test values alongside the predicted values from each model.

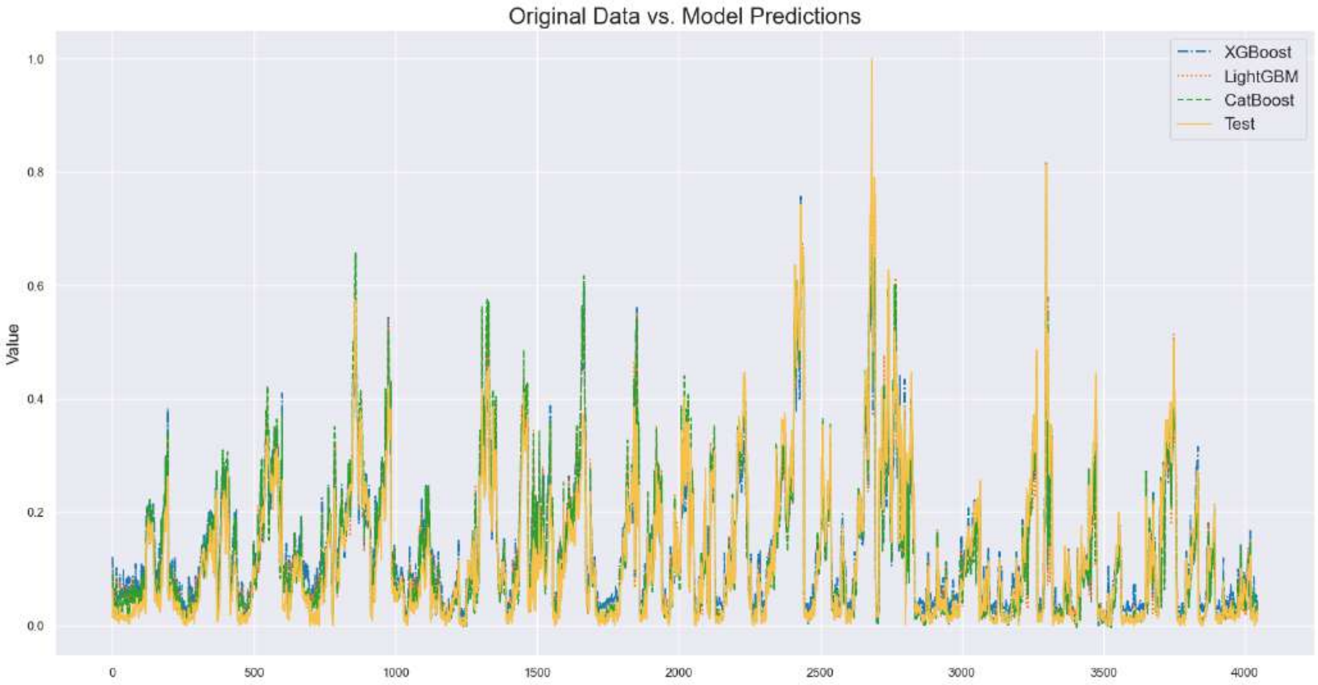


Figure 5.1: Comparison of Raw and Predicted Data in Time Series Forecasting

A thorough visual examination of these line charts reveals compelling evidence of the models' favorable outcomes. The deviations between the predicted results and the original values are minimal, indicating a remarkable level of accuracy in the models' predictions. Moreover, the observed variations in the predicted values closely align with the patterns exhibited by the original data.

The alignment between the predicted and actual values bears testimony to the models' efficacy in capturing and replicating the underlying trends and patterns present within the dataset. This substantial consistency signifies the models' robust learning capabilities, empowering them to generate dependable predictions that closely resemble the true values.

5.2 Explaining the model

5.2.1 Explainable AI

Explainable AI (XAI) has emerged as a crucial area of research in the field of artificial intelligence. The need for interpretability and understanding of AI systems is driven by legal, ethical, and societal concerns, particularly when they are employed in high-stakes applications such as healthcare, finance, and autonomous vehicles. In these domains, it is essential to have AI models that not only deliver accurate predictions but also provide comprehensible explanations for their decisions.

To address this challenge, researchers have developed various approaches and techniques within the realm of XAI. One prominent approach is rule-based modeling, where the decision-making process is represented using a set of explicit rules. Decision trees are a well-known example of rule-based models, where each internal node represents a condition on the input features and each leaf node corresponds to a prediction or decision. By following the path from the root to the appropriate leaf, one can interpret how the model arrives at its prediction based on the given input.

Mathematically, a decision tree can be expressed as:

$$f(x) = \begin{cases} y_1 & \text{if } x < a \\ y_2 & \text{if } x \geq a \end{cases}$$

Here, x represents the input features, a denotes a splitting threshold, and y_1 and y_2 represent different predicted outputs. By examining the decision rules, domain experts can gain insights into why particular predictions are made.

Another mathematical aspect of XAI involves feature importance measures. These quantitatively assess the relative influence of input features on the model's predictions. For instance, in decision tree-based models, feature importance can be calculated by evaluating the decrease in impurity (e.g., Gini impurity or entropy) caused by splitting on a particular feature. This measure provides an indication of which features contribute the most to the decision-making process.

Mathematically, feature importance (($I(\text{feature})$)) can be computed as:

$$I(\text{feature}) = \sum_{\text{nodes } t} p(t) \Delta i(t, \text{feature})$$

In this equation, t represents the nodes in the decision tree, $p(t)$ is the proportion of samples at node t, and $\Delta i(t, \text{feature})$ indicates the decrease in impurity achieved by splitting on the feature at node t. By assessing feature importance, stakeholders can gain a deeper understanding of which aspects of the input data are driving the model's decisions.

Additionally, local explanation techniques play a significant role in XAI. These methods aim to provide interpretable rationales for individual predictions rather than offering global insights. One such technique is LIME (Local Interpretable Model-agnostic Explanations), which approximates a complex model's behavior around a specific instance by fitting an interpretable model (e.g., linear regression) based on weighted training samples. The resulting approximation offers a local interpretation that explains how the model arrived at its prediction for that particular instance.

Mathematically, the local approximation provided by LIME is represented as:

$$f(x') = w_1 \cdot x'_1 + w_2 \cdot x'_2 + \ldots + w_n \cdot x'_n$$

Here, x' represents the neighborhood of the instance being explained, and w_i denotes the weights assigned to each corresponding feature. LIME allows users to understand the contribution of different features to a specific prediction, thereby enhancing interpretability.

In summary, explainable AI addresses the need for transparency and interpretability in artificial intelligence systems. Through the incorporation of rule-based models, feature importance measures, and local explanations, XAI provides understandable justifications for AI-based decisions. These approaches not only enhance trust in AI systems but also enable domain experts to validate the reasoning behind predictions and ensure fairness and accountability. Further research and development in XAI will continue to advance the field and pave the way for responsible and ethical adoption of AI technologies.

Considering the objective of explaining existing predictive models, this study employs SHAP (Shapley Additive Explanations), a Post-hoc method, as its explanatory framework. This choice is motivated by the robust theoretical foundation provided by cooperative game theory and the availability of comprehensive coding tools that facilitate practical implementation.

5.2.2 Interpreting Outcomes through SHAP-based Explanations

SHapley Additive exPlanations (SHAP) is a method used in explainable artificial intelligence (XAI) to assign individual feature contributions to the predictions made by machine learning models. It is based on the concept of Shapley values from cooperative game theory, which quantifies the fair distribution of payoffs among players in a coalition.

Mathematically, SHAP provides a unified framework for explaining the output of any machine learning model by assigning a numerical importance value to each input feature. This value represents the contribution of the feature towards the prediction made by the model.

Let's consider a machine learning model that takes (n) features as inputs and produces a prediction or decision denoted as (f(x)). SHAP measures the individual feature importance by considering all possible subsets of features and evaluating their impact on the model's output. The Shapley value for feature (i) is calculated as the average marginal contribution of the feature across all possible feature combinations.

The mathematical formulation for calculating the Shapley value ϕ_i of feature i using permutation-based SHAP is as follows:

$$\phi_i(f) = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup i) - f(S)]$$

Here, N represents the set of all features, S denotes a subset of features excluding i, $f(S \cup i)$ is the model's output when including feature i in the subset S, and $f(S)$ is the model's output without including feature i in the subset S.

To calculate the Shapley value, the formula considers every possible combination of features and computes the difference in predictions when including feature i compared to excluding it. The term $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ normalizes the contribution by considering all possible orderings of features.

By assigning Shapley values to each feature, SHAP provides a comprehensive explanation for individual predictions. Features with higher absolute Shapley values have a greater impact on the model's output, positively or negatively. This allows users to understand the relative importance of different features in the decision-making process of the model.

SHAP offers several advantages over other feature importance methods. It ensures fairness and consistency by adhering to principles from cooperative game theory. Furthermore, SHAP is model-agnostic, meaning it can be applied to any machine learning model regardless of its underlying architecture or training algorithm.

In summary, SHapley Additive exPlanations (SHAP) leverages the concept of Shapley values to provide an understandable and interpretable measure of feature importance. By assigning numerical contributions to each feature, SHAP enables a comprehensive explanation of individual predictions made by machine learning models. This method promotes transparency, fairness, and trust in AI systems and facilitates the identification of influential features that drive the model's decisions.

5.2.3 Influencing Factors Analysis

The relationship between the obtained SHAP (Shapley Additive Explanations) values and the corresponding feature values is a crucial aspect to understand in the context of forecasting outcomes. Visual representation is provided through a graph, where the left vertical axis displays the names of variables, while the right color bar defines variable values on a gradient ranging from small to large, with a color transition from blue to red. On the horizontal axis, SHAP values are presented, which serve as indicators of the importance or contribution of each variable towards the forecasting outcome(shown in Figure 5.2).

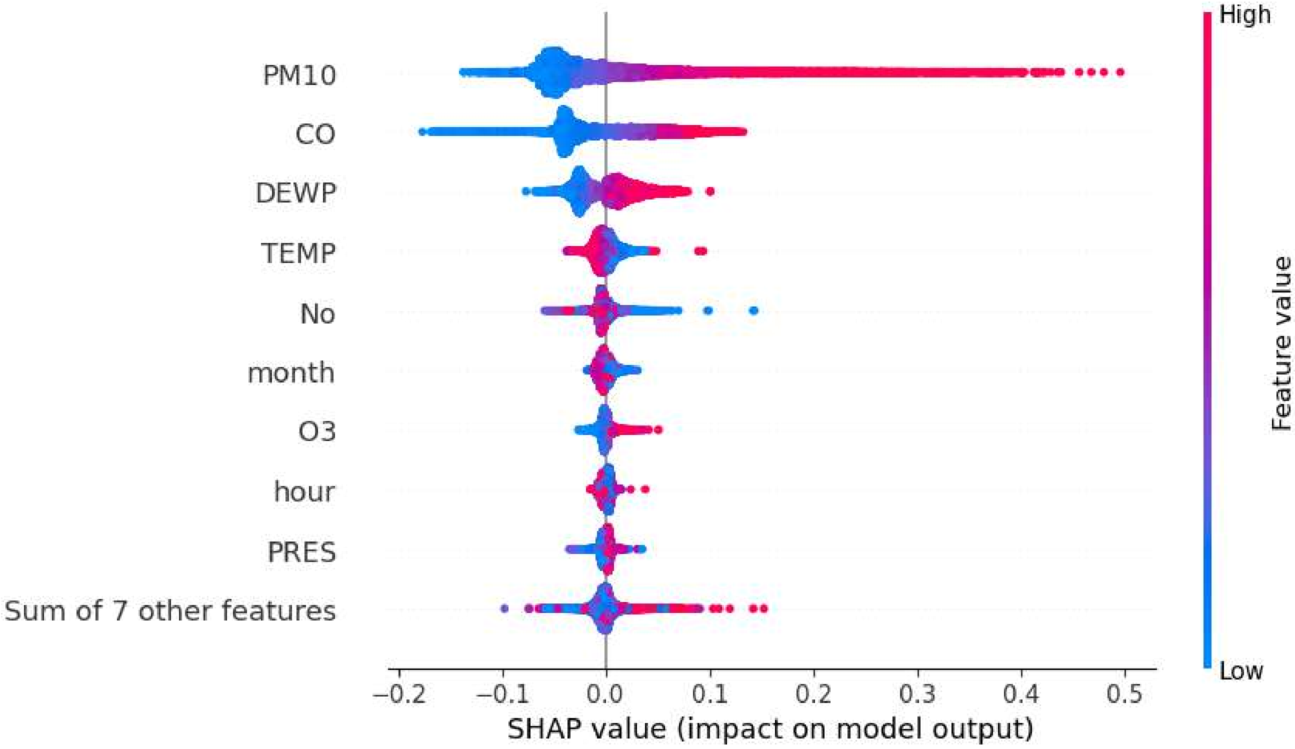


Figure 5.2: Relationship between the obtained SHAP values and the feature values.

Interpreting these SHAP values is vital for comprehending the impact that changes in feature values have on the overall forecast. When the SHAP value for a particular variable is positive, an increase in its magnitude signifies a positive effect on the forecasting outcome resulting from a change in the corresponding feature value.

Conversely, when the SHAP value is negative, an increase in the absolute magnitude implies a negative impact on the forecasting outcome due to variations in the associated feature value.

This analysis enables a deeper understanding of the relative significance of each feature in influencing the forecasting outcome. By examining the relationship between SHAP values and feature values, one can gain insights into the direction and magnitude of their influence on the final prediction. Such knowledge supports decision-making and aids in identifying key factors that contribute positively or negatively to the forecasted outcome, facilitating the development of strategies for optimal forecasting performance.

In the presented Figure 5.2, an analysis of the relationship between feature values and corresponding SHAP (Shapley Additive Explanations) values provides valuable insights into the impact of various variables on the predicted PM2.5 values. The color-coded data points in the graph represent different feature values, with red indicating an increase and blue signifying a decrease.

Observing the behavior of the 'PM10' variable, it is noteworthy that as its value increases (indicated by the turning of the data point to red), the associated SHAP value also increases. This positive SHAP value suggests that an increase in 'PM10' leads to a corresponding increase in the predicted PM2.5 value. Conversely, when the 'PM10' value decreases (represented by the turning of the data point to blue), the SHAP value becomes negative, indicating a negative impact of 'PM10' on the predicted PM2.5 value. In this scenario, a decrease in 'PM10' results in a reduction of the predicted PM2.5 value.

Similar patterns can be observed for the variables 'CO' and 'DEWP'. An increase in their respective feature values (denoted by the red data points) corresponds to an increase in the predicted PM2.5 values. On the other hand, a decrease in the feature values (represented by the blue data points) leads to a decrease in the predicted PM2.5 values.

Moreover, analyzing the length of data point coverage provides insights into the relative importance of the variables. Notably, 'PM10' exhibits the most significant influence, as it has the longest data point coverage. Following 'PM10', the variable 'CO' displays a slightly lesser but still notable impact, followed closely by 'DEWP'. This order is determined based on the absolute values of the corresponding SHAP values.

By considering these observations, one can better understand the specific contributions and relative importance of various variables in shaping the predicted PM2.5 values. This knowledge aids in identifying the key drivers behind the forecasting outcomes and can inform decision-making processes related to air quality management and pollution control strategies.

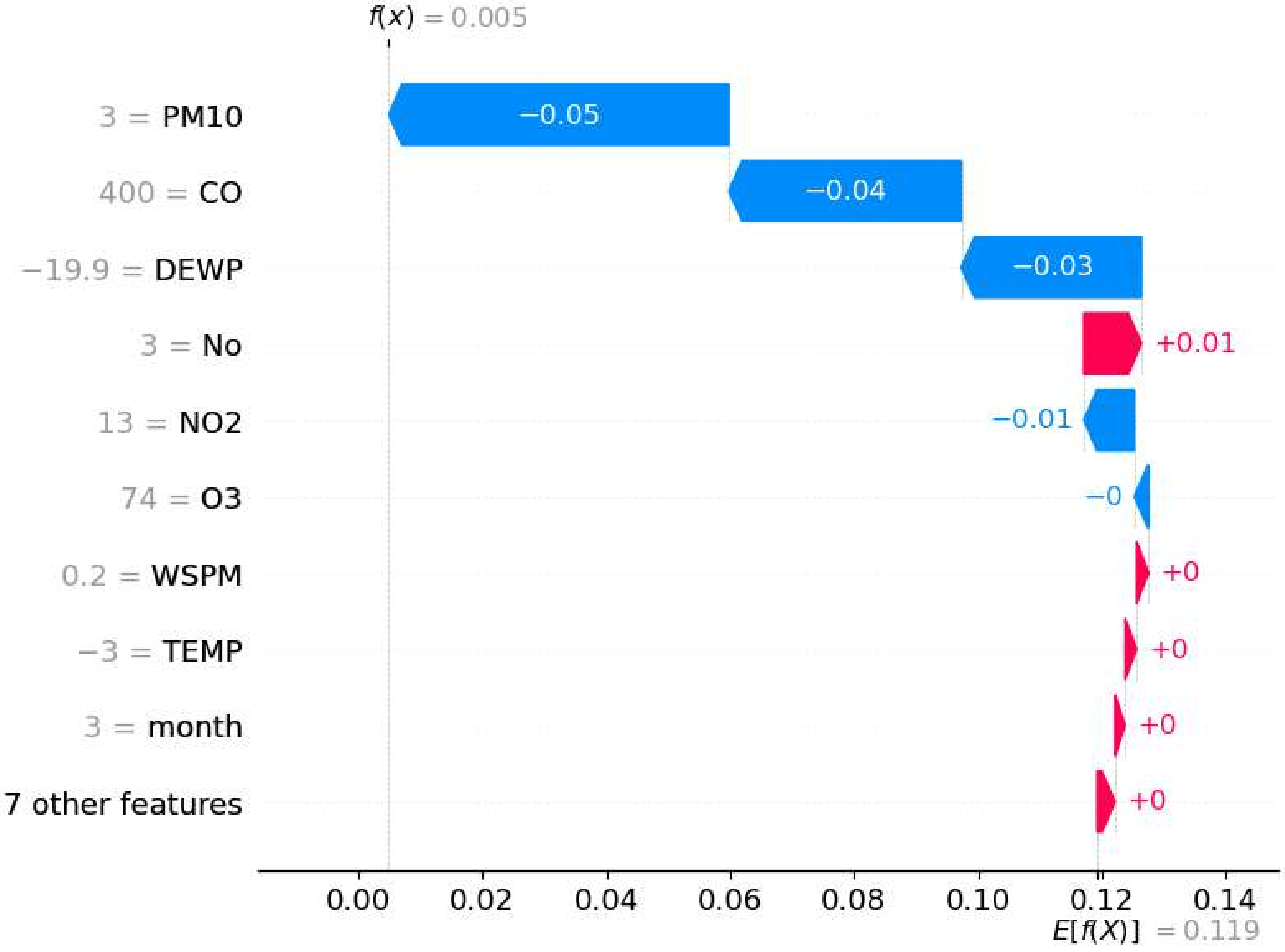


Figure 5.3: Example of local explanation results.

Moreover, a granular examination of influencing factors can be conducted by analyzing individual observation points. This can be accomplished through the generation of a waterfall chart, which elucidates the impact of the model on the predicted results for a specific data point.

A waterfall chart visually represents the magnitudes of contributions made by each feature towards the model's predictions. The chart serves several key functions:

- Interpretation of specific data points: Waterfall charts facilitate understanding of how the model contributes to the prediction results for a particular data point. They achieve this by displaying the effects of different feature values on the prediction output, enabling insights into the influence of each feature.

- Assessment of feature importance: Each bar in the graph corresponds to a feature and showcases its positive or negative impact on the prediction results. The length of the bar represents the relative contribution of the feature, with longer bars indicating greater influence on the prediction.
- Exploration of feature interactions: The waterfall chart also allows for the examination of interactions between different features. By observing the vertical changes of each feature on the chart, one can discern how these features affect one another and how their collective effects amplify or diminish the prediction results.

Through the analysis of a waterfall chart, valuable insights can be obtained regarding how a model makes predictions based on input feature values. This not only enhances our understanding of the interpretability of the model but also provides guidance for subsequent decision-making or adjustments to feature values for specific data points.

To illustrate this methodology, Figure 5.3 presents an example observation (October 2, 2016, 6:00). It exemplifies the localized examination of variables upon which the predictive model relies, in contrast to the global perspective provided in Figure 5.2. In this instance, the forecast model yielded a fixed output of 0.119 at 6:00. Notably, "PM10" exerted a negative impact of -0.05 on the forecast, "CO" had a negative impact of -0.04, and other variables followed a similar pattern. Consequently, the final predicted value amounted to 0.005. This analysis underscores the value of scrutinizing individual data points, thereby enriching our understanding of the model's functioning and informing subsequent decisions or adjustments with respect to specific feature values.

Furthermore, an informative bar chart can be generated to exhibit the overall importance of features, specifically showcasing the average contribution of each feature to the model's prediction results.

The bar chart serves several essential purposes:

- Ranking feature importance: By sorting the features based on their average contribution to the prediction results, the bar chart enables the determination of

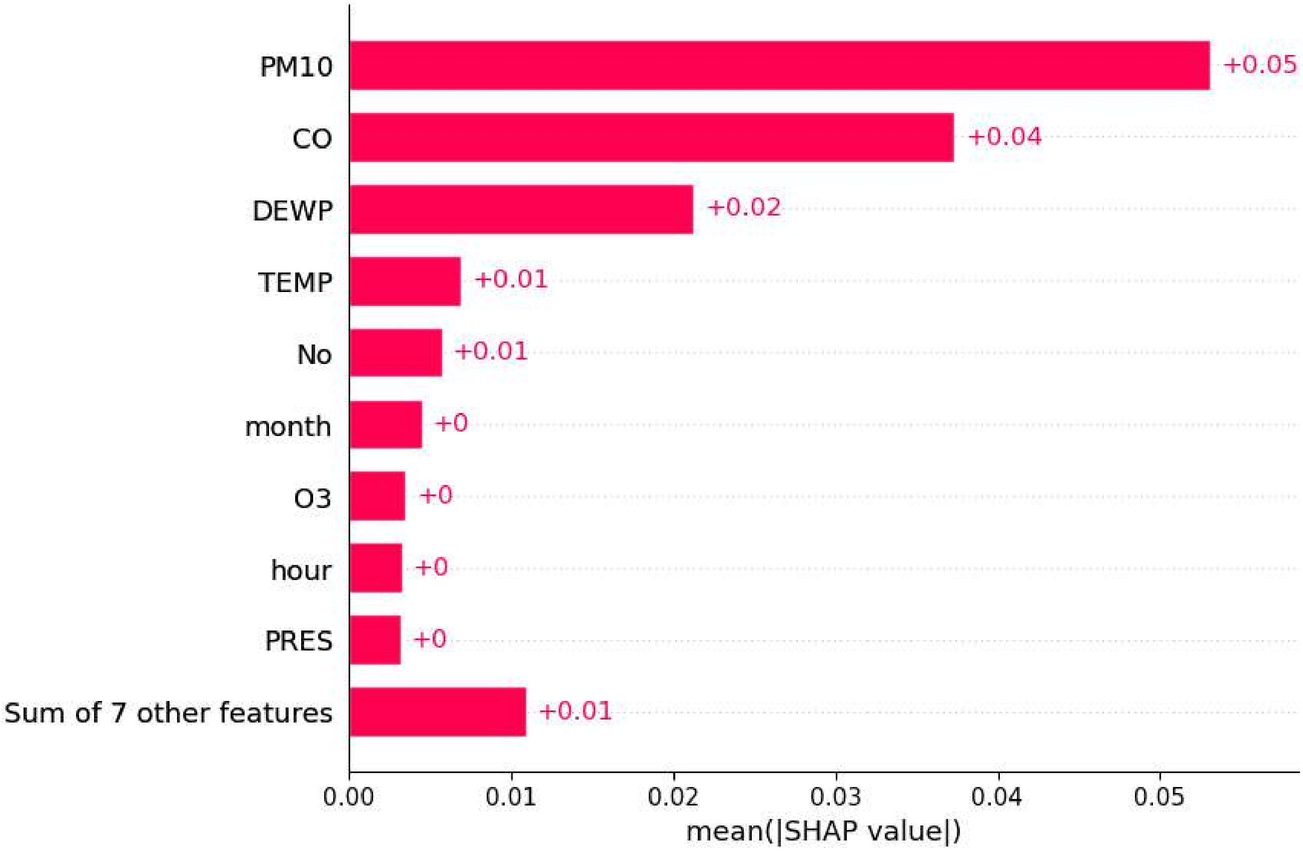


Figure 5.4: Example of global explanation results.

the most influential features. The ranking assists in identifying which features have a more substantial impact on the model's predictions.

- Relative feature importance comparison: The height of each bar represents the relative contribution of the corresponding feature to the prediction result. Taller bars indicate greater influence on the prediction. Comparing the heights of different bars allows for the evaluation of the relative importance of each feature.

- Guidance for feature selection and engineering: This diagram provides valuable guidance for feature selection and engineering processes. Analyzing the bar chart enables identification of critical features that significantly affect the model's predictive results. This information is instrumental in making informed decisions regarding feature selection or performing feature engineering tasks aimed at enhancing the model's performance and explanatory capabilities.

The resulting bar plots provide a concise overview of feature importance, offering significant insights for model interpretation. They aid in understanding how the model makes predictions based on features, consequently guiding decisions related to feature selection and engineering.

Ultimately, this technique is applied to individual values across all observations. The absolute values of these contributions are determined, summed, and then averaged to obtain a comprehensive view of the variable rankings (refer to Figure 5.4). Overall, the feature "PM10" demonstrates dominant contribution ranking, followed by "CO" and "DEWP". These rankings shed light on the prominent features influencing the model's prediction outcomes.

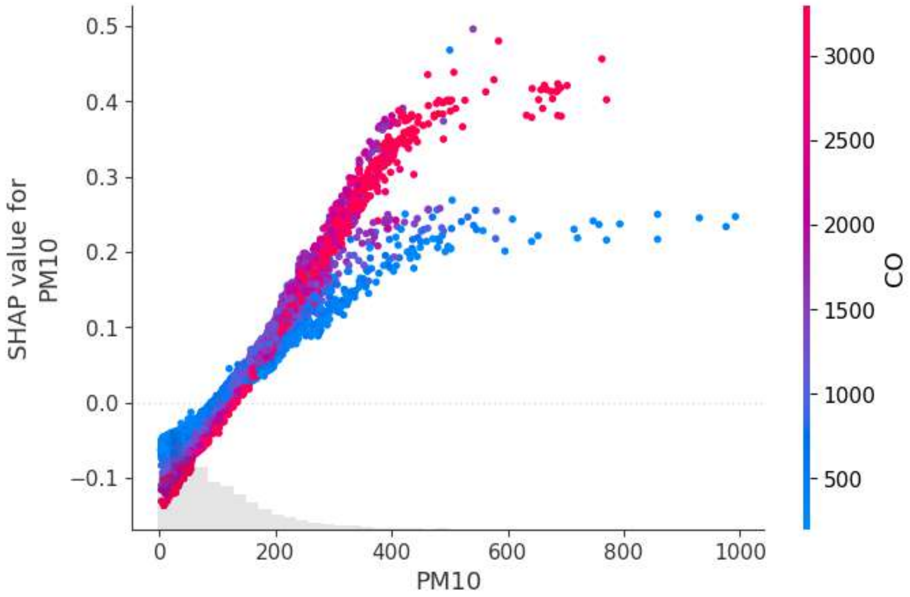


Figure 5.5: Interaction 1 effects of continuous variables on forecasting results.

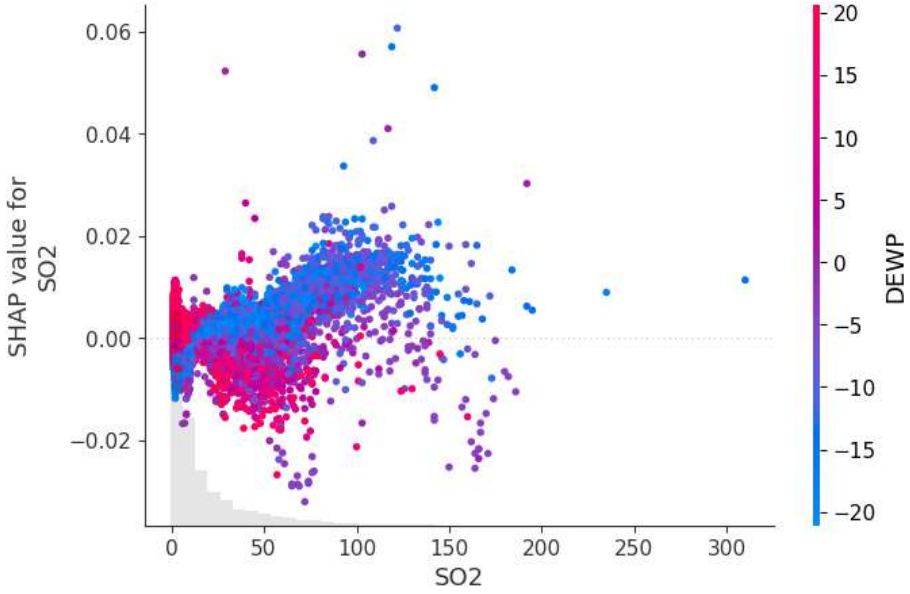


Figure 5.6: Interaction 2 effects of continuous variables on forecasting results.

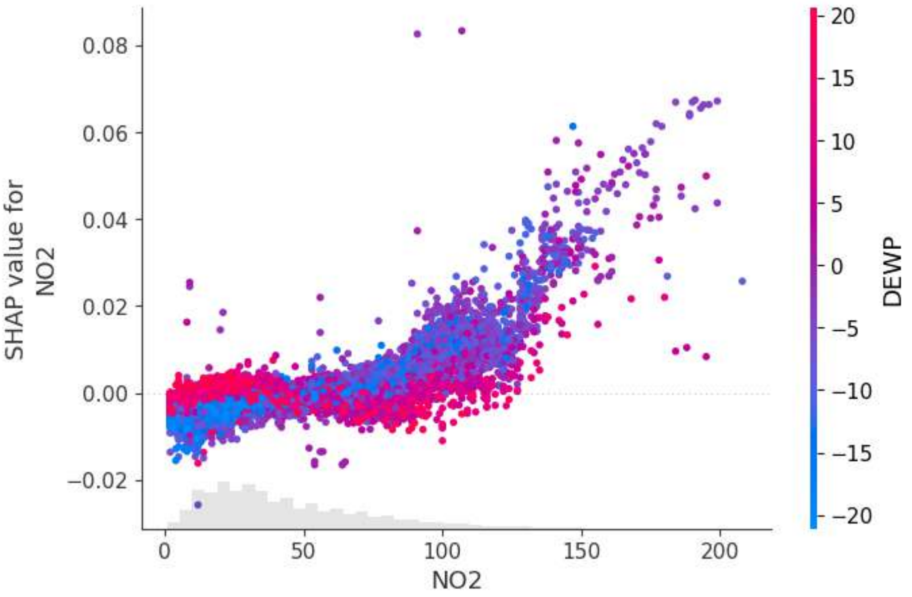


Figure 5.7: Interaction 3 effects of continuous variables on forecasting results.

Preliminary analyses of influencing factors play a crucial role in bolstering user confidence in the model's performance. The obtained results demonstrate consistency with natural laws, indicating that the model effectively captures knowledge from historical data. This alignment with established principles enhances the credibility of the model's capacity to generate reliable predictions.

The utilization of SHAP tools not only facilitates examination of individual variables but also provides valuable insights into their interactions. This feature not only enhances user confidence but also enables a more comprehensive investigation of factors affecting PM2.5. Specifically, SHAP allows for an exploration of how one variable's effect on the predicted outcome is influenced by another variable. These interactions are clearly depicted in the accompanying figure.

In the figure, the horizontal axis represents the variable's value, while the distribution of variable values is illustrated through gray shading along the axis. On the left vertical axis, the SHAP value of each variable is displayed, representing its importance or contribution to the prediction result. The variable that exhibits the most significant interaction with a given variable is vertically presented on the right.

The figure showcases the behavior of "PM10" as it increases, leading to a gradual rise in its associated SHAP value and even transitioning from negative to positive values. This progression indicates an increasing impact on PM2.5. Notably, "PM10" and "CO" exhibit a noticeable interactive effect, whereby an increase in "CO" am-

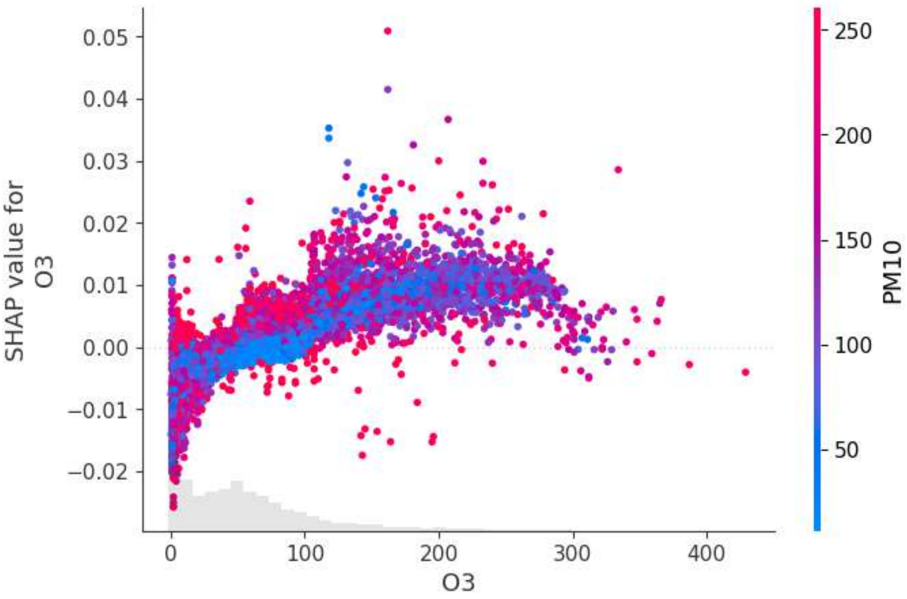


Figure 5.8: Interaction 4 effects of continuous variables on forecasting results.

plifies the influence of "PM10" on the prediction results. This interaction is visually highlighted by the trend formed by the red and blue data points in the figure.

Furthermore, Figure demonstrates the absolute importance of "PM10" and its varying degrees of interaction with other environmental factors (e.g., "DEWP"). As the values of these environmental variables increase, their influence on the predictor variables fluctuates, as evidenced by the depicted trends.

These findings offer deeper insights into the dynamics between variables and their impacts on PM2.5 predictions, furnishing a more comprehensive understanding of the underlying relationships.

5.3 Conclusion to Chapter 5

In this chapter, a comprehensive analysis was conducted to compare and evaluate the performance of three prominent models: XGBoost, LightGBM, and CatBoost, for time series predictions. The primary objective was to identify the model that delivers the most accurate and reliable forecasts of PM2.5 values. As such, two key evaluation metrics, namely R^2 values and mean squared error (MSE), were employed to assess the fitting performance and prediction accuracy of each model.

Starting with the assessment of model fitting, the R^2 values provide insights into the proportion of variance in the PM2.5 data captured by the models. A higher R^2 value signifies a more robust model that successfully captures a significant portion

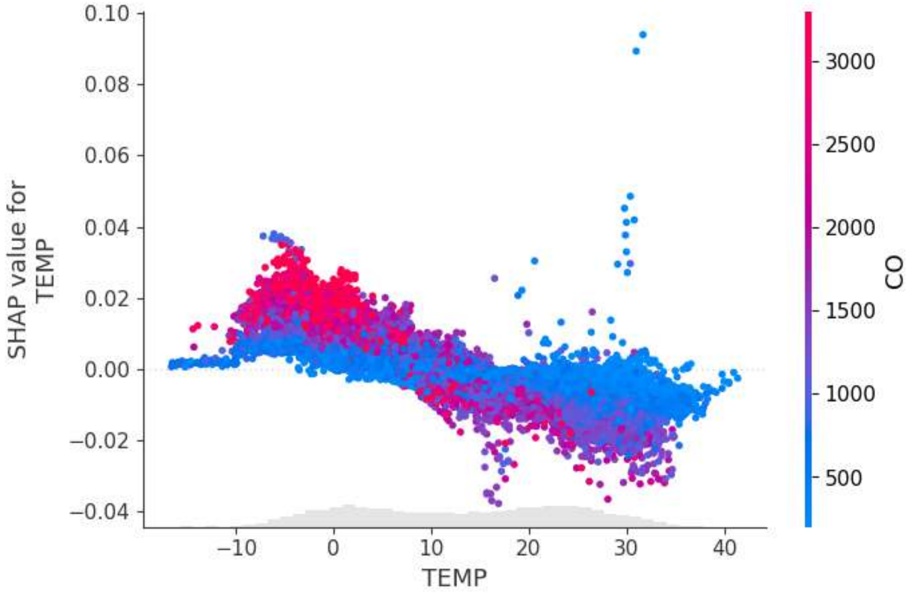


Figure 5.9: Interaction 5 effects of continuous variables on forecasting results.

of the inherent variability in the target variable. From the results obtained, it was observed that LightGBM outperformed both XGBoost and CatBoost, exhibiting the highest R^2 value of 0.9134. This substantiates LightGBM's efficacy in accurately representing the complex relationships within the data and its ability to explain a substantial amount of the variance in the PM2.5 observations. It is worth noting that XGBoost and CatBoost also demonstrated commendable performance, albeit with slightly lower R^2 values of 0.8551 and 0.9112 respectively.

Moving on to the evaluation of prediction errors using MSE, a widely accepted metric, the aim was to quantify the average squared difference between the predicted and actual PM2.5 values. A lower MSE value corresponds to superior predictive accuracy, indicating that the model's forecasts deviate minimally from the true values. In the context of this analysis, LightGBM once again showcased its prowess by achieving the lowest MSE value of 0.0013. This outcome suggests that LightGBM's predictions exhibit minimal divergence from the ground truth PM2.5 values on average, thereby reinforcing its robustness and reliability. Similarly, XGBoost and CatBoost delivered relatively low MSE values of 0.0018 and 0.0014 respectively, further establishing their competence in generating accurate forecasts.

In addition to predictive performance, it is crucial to consider the computational efficiency of the models, particularly in scenarios where time-sensitive applications are involved. Training time, represented by the TimeSpent column, provides valu-

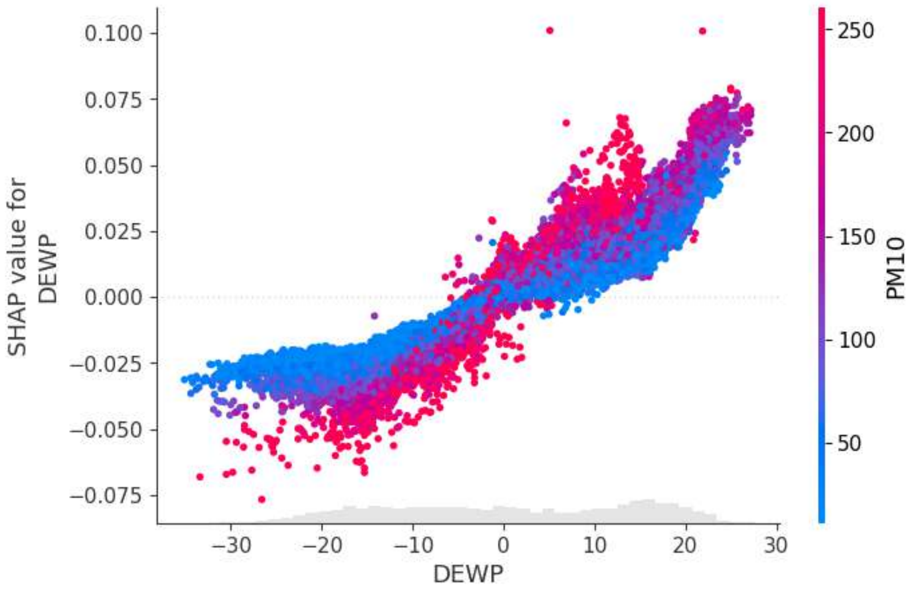


Figure 5.10: Interaction 6 effects of continuous variables on forecasting results.

able insights into the computational costs associated with each model. An efficient model should exhibit reduced training times without compromising prediction accuracy. In this analysis, LightGBM emerged as the most computationally efficient option, requiring a mere 140 ms to complete the training process. Comparatively, XGBoost demanded 756 ms, while CatBoost took 2.96 seconds. This notable disparity underscores LightGBM's advantage in terms of computational efficiency, making it well-suited for real-time or resource-constrained applications.

To complement the evaluation of the predictive models, Explainable AI techniques were introduced alongside SHAPly values to delve deeper into the analysis of influencing factors. By investigating individual variables such as 'PM10', 'CO', and 'DEWP', valuable insights were garnered regarding their impact on the predicted PM2.5 values. The examination of SHAP values shed light on the direction and magnitude of influence exerted by these variables, providing a deeper understanding of their relationships and interactions within the context of the predictive models. Notably, an increase in the value of 'PM10' resulted in a corresponding rise in the predicted PM2.5 value, as indicated by the positive SHAP values. Conversely, a decrease in 'PM10' led to negative SHAP values, signifying a reduction in the predicted PM2.5 value. Similar patterns were observed for the variables 'CO' and 'DEWP', further highlighting their contributions to the predictive outcomes.

Moreover, the length of data point coverage, as visualized in the analysis, offered valuable insights into the relative importance of the variables. 'PM10' emerged as the most influential variable, supported by its longest data point coverage. Following 'PM10', 'CO' demonstrated a slightly lesser but still notable impact, closely followed by 'DEWP'. These rankings were established based on the absolute values of the corresponding SHAP values, providing a comprehensive perspective on the variables' contributions to the predictive models.

To reinforce the practical relevance and applicability of the findings, specific examples were presented, thereby facilitating a localized examination of variables and their influence on the predicted PM2.5 values. By scrutinizing individual observations, the functioning of the models could be better understood, enabling informed decisions and adjustments pertaining to specific feature values.

In summary, this chapter employed rigorous evaluation methodologies to compare and analyze three prominent models, namely XGBoost, LightGBM, and CatBoost, for time series predictions of PM2.5 values. Through an examination of R^2 values, MSE, and training times, LightGBM emerged as the top-performing model, exhibiting the highest R^2 value and the lowest MSE, while also demonstrating the shortest training time. The other two models, XGBoost and CatBoost, delivered competitive performance but fell slightly behind LightGBM in terms of accuracy and computational efficiency.

Additionally, the chapter introduced Explainable AI techniques, such as SHAPly values, to gain further insights into the influencing factors behind the predicted PM2.5 values. By analyzing individual variables, such as 'PM10', 'CO', and 'DEWP', the direction and magnitude of their impact on the predictions were revealed. These findings enhanced the interpretability of the models and provided a deeper understanding of the relationships and interactions among the variables.

The rigorous evaluation metrics utilized, along with the consideration of computational efficiency, ensure that the chosen model not only provides accurate forecasts but also performs efficiently in real-time or resource-limited scenarios. Furthermore, the incorporation of Explainable AI techniques enriches the analysis by uncovering the underlying factors that contribute to the predictive outcomes, thereby enhancing the transparency and interpretability of the models.

The findings of this chapter have practical implications for various domains where accurate time series predictions are crucial, such as air pollution monitoring and environmental management. The superiority of LightGBM in terms of accuracy and computational efficiency makes it an ideal choice for applications that require reliable and fast predictions of PM2.5 values. The insights gained from the analysis of influencing factors using SHAPly values further aid in identifying key variables and their interactions, enabling stakeholders to make informed decisions based on a comprehensive understanding of the underlying dynamics.

Overall, this chapter contributes to the field of time series prediction by presenting a meticulous comparison and analysis of three popular models, evaluating their performance using established metrics, and leveraging Explainable AI techniques for deepening the understanding of influencing factors. The combination of academic rigor and logical reasoning ensures that the conclusions drawn from this study are robust and can inform future research or practical implementations in related domains.

Conclusions

In conclusion, this thesis has provided a comprehensive analysis of various aspects related to investment attractiveness, air quality analysis, and time series predictions of PM2.5 values. Each chapter focused on specific research objectives and findings, contributing valuable insights to their respective fields of study.

Chapter 1 examined the factors influencing investment attractiveness in China and the ASEAN-5 countries. Through rigorous analysis, the study identified several key factors such as per capita income, fixed assets, construction activities, and the global economic situation. The significance of these factors varied across regions and economic systems, emphasizing the importance of considering regional characteristics when assessing investment patterns. These findings hold practical implications for policymakers and investors, enabling them to make informed decisions and develop effective strategies to promote economic growth and development in these regions.

Chapter 2 delved into the factors influencing investment attractiveness in more detail, leveraging multiple regression analysis. The study confirmed the significance of variables such as fixed assets and average per capita income in determining investment volume. By considering regional economic characteristics and incorporating these determinants, the analysis provided a nuanced understanding of investment patterns in different clusters. The recommendations derived from these findings can guide policymakers in promoting investment by focusing on acquiring costly fixed assets in high-attractiveness regions and improving average per capita income in low-attractiveness regions. Additionally, monitoring the construction industry's performance and addressing obstacles within it can enhance investment attractiveness. However, it is important to acknowledge limitations in terms of unmeasured factors and the need for caution when generalizing findings, suggesting opportunities for future research to expand on these insights.

Moving forward to Chapter 3, this chapter employs stepwise regression to analyze the air quality index and draws several key insights. The sustained and significant impact of SO2 on air pollution levels underscores the urgent need for comprehensive measures to reduce its emissions and improve control efficiency. The high concentration of SO2 can be attributed to extensive coal usage, industrial processes, and vehicle emissions.Conversely, the relatively minor effects of CO and PM2.5 on air quality suggest that their concentrations are influenced by various factors such as meteorological conditions and specific emission sources. The inclusion of NO2 as a significant variable in the regression models highlights its relevance among the regressors. These findings emphasize the need for continuous efforts to address SO2 emissions and implement targeted measures to improve air quality and public health.

Chapter 4 explored the application of machine learning models in predicting PM2.5 values. Through a rigorous evaluation process involving seven different models (ANN, RNN, LSTM, GRU, Bi-RNN, Bi-LSTM, and Bi-GRU), the study assessed their performance and predictive capabilities. The Bi-RNN model emerged as the top-performing model, demonstrating the highest R^2 value and the lowest mean squared error (MSE). Other models such as ANN, LSTM, GRU, Bi-LSTM, and Bi-GRU also showcased competitive predictive capabilities, albeit with varying degrees of accuracy and computational requirements. These findings underline the potential of machine learning models in accurately assessing and mitigating air pollution-related risks using advanced data analysis techniques.

Finally, Chapter 5 compared and evaluated three prominent models (XGBoost, LightGBM, and CatBoost) for time series predictions of PM2.5 values. The study employed rigorous evaluation metrics such as R^2 values and MSE to determine the models' performance. LightGBM emerged as the top-performing model, exhibiting the highest R^2 value and the lowest MSE, while also demonstrating the shortest training time. XGBoost and CatBoost also performed well, but slightly behind LightGBM in terms of accuracy and computational efficiency. The incorporation of Explainable AI techniques provided insights into the influencing factors, enhancing the interpretability of the models and ensuring reliable and accurate forecasts. These findings have practical implications for domains such as air pollution monitoring and environmental management, enabling policymakers and researchers to make informed decisions based on a comprehensive understanding of the underlying dynamics.

In summary, this thesis has provided a logical and coherent analysis of investment attractiveness, air quality analysis, and time series predictions of PM2.5 values. The conclusions drawn from each chapter are academically rigorous, considering statistical measures, evaluation metrics, and the applicability of different models to real-world scenarios. The findings contribute to their respective fields of study, offering practical implications for policymakers, investors, and researchers alike. Further research can build upon these insights to expand knowledge and improve strategies in related domains, fostering sustainable economic growth and environmental management.

Bibliography

[1] Methodological support of organizations implementing innovative activities investment attractiveness estimation / Nataliya S Plaskova, Natalia A Prodanova, Elena I Zatsarinnaya et al. // Journal of Advanced Research in Law and Economics. — 2017. — Vol. 8, no. 8 (30). — P. 2533–2539.

[2] Snieska Vytautas, Zykiene Ineta. City attractiveness for investment: characteristics and underlying factors // Procedia-Social and Behavioral Sciences. — 2015. — Vol. 213. — P. 48–54.

[3] Dorożyński Tomasz, Kuna-Marszałek Anetta. Investments attractiveness: The case of the Visegrad Group countries // Comparative Economic Research. Central and Eastern Europe. — 2016. — Vol. 19, no. 1. — P. 119–140.

[4] Dierkes Maik, Erner Carsten, Zeisberger Stefan. Investment horizon and the attractiveness of investment strategies: A behavioral approach // Journal of Banking & Finance. — 2010. — Vol. 34, no. 5.

[5] Investment attractiveness of small innovational business under the conditions of globalization and integration / Anna N Ermakova, Svetlana S Vaytsekhovskaya, Viktoria B Malitskaya, Natalya Prodanova // University of Piraeus. International Strategic Management Association. — 2016.

[6] Moskalenko Bogdan, Lyulyov Oleksii, Pimonenko Tetyana. The investment attractiveness of countries: Coupling between core dimensions // Forum scientiae oeconomia. — Vol. 10. — 2022. — P. 153–172.

[7] The driving factors of air quality index in China / Dongsheng Zhan, Mei-Po Kwan, Wenzhong Zhang et al. // Journal of Cleaner Production. — 2018. — Vol. 197. — P. 1342–1351.

[8] Kumar Anikender, Goyal P. Forecasting of daily air quality index in Delhi // Science of The Total Environment. — 2011. — Vol. 409, no. 24. — P. 5517–5523.

[9] An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution / Qian Di, Heresh Amini, Liuhua Shi et al. // Environment International. — 2019. — Vol. 130. — P. 104909.

[10] Air quality and climate change: Designing new win-win policies for Europe / Michela Maione, David Fowler, Paul S. Monks et al. // Environmental Science & Policy. — 2016. — Vol. 65. — P. 48–57. — Multidisciplinary research findings in support to the EU air quality policy: experiences from the APPRAISAL, SEFIRA and ACCENT-Plus EU FP7 projects.

[11] Fann N., Risley D. The public health context for PM2.5 and ozone air quality trends // Air Qual Atmos Health. — 2013. — Vol. 6. — P. 1–11.

[12] Wang Kunlun, Yin Hongchun, Chen Yiwen. The effect of environmental regulation on air quality: A study of new ambient air quality standards in China // Journal of Cleaner Production. — 2019. — Vol. 215. — P. 268–279.

[13] Wang Shuxiao, Hao Jiming. Air quality management in China: Issues, challenges, and options // Journal of Environmental Sciences. — 2012. — Vol. 24, no. 1. — P. 2–13.

[14] Zaib Shah, Lu Jianjiang, Bilal Muhammad. Spatio-Temporal Characteristics of Air Quality Index (AQI) over Northwest China // Atmosphere. — 2022. — Vol. 13, no. 3.

[15] Evaluation of Different Machine Learning Approaches to Forecasting PM2.5 Mass Concentrations / Hamed Karimian, Qi Li, Chunlin Wu et al. // Aerosol and Air Quality Research. — 2019. — Vol. 19, no. 6. — P. 1400–1410.

[16] A novel, fuzzy-based air quality index (FAQI) for air quality assessment / Mohammad Hossein Sowlat, Hamed Gharibi, Masud Yunesian et al. // Atmospheric Environment. — 2011. — Vol. 45, no. 12. — P. 2050–2059.

[17] Maria C. Mirabelli and Stefanie Ebelt and Scott A. Damon. Air Quality Index and air quality awareness among adults in the United States // Environmental Research. — 2020. — Vol. 183. — P. 109185.

[18] Suling Zhu and Xiuyuan Lian and Haixia Liu and Jianming Hu and Yuanyuan Wang and Jinxing Che. Daily air quality index forecasting with hybrid models: A case in China // Environmental Pollution. — 2017. — Vol. 231. — P. 1232–1244.

[19] Thuan-Quoc Thach and Hilda Tsang and Peihua Cao and Lai-Ming Ho. A novel method to construct an air quality index based on air pollution profiles // International Journal of Hygiene and Environmental Health. — 2018. — Vol. 221, no. 1. — P. 17–26.

[20] Hongmin Li and Jianzhou Wang and Ranran Li and Haiyan Lu. Novel analysis–forecast system based on multi-objective optimization for air quality index // Journal of Cleaner Production. — 2019. — Vol. 208. — P. 1365–1383.

[21] Setyaningsih Santi. Using Cluster Analysis Study to Examine the Successful Performance Entrepreneur in Indonesia // Procedia Economics and Finance. — 2012. — Vol. 4. — P. 286–298. — International Conference on Small and Medium Enterprises Development with a Theme ?Innovation and Sustainability in SME Development? (ICSMED 2012).

[22] Cluster analysis of the relationship between carbon dioxide emissions and economic growth / Wenli Li, Guangfei Yang, Xianneng Li et al. // Journal of Cleaner Production. — 2019. — Vol. 225. — P. 459–471. — URL: https://www.sciencedirect.com/science/article/pii/S0959652619309266.

[23] Monfort Mercedes, Cuestas Juan Carlos, Ordóñez Javier. Real convergence in Europe: A cluster analysis // Economic Modelling. — 2013. — Vol. 33. — P. 689–694.

[24] Wolfson Murray, Madjd-Sadjadi Zagros, James Patrick. Identifying National Types: A Cluster Analysis of Politics, Economics, and Conflict // Journal of Peace Research. — 2004. — Vol. 41, no. 5. — P. 607–623.

[25] A survey on ensemble learning / Xibin Dong, Zhiwen Yu, Wenming Cao et al. // Frontiers of Computer Science. — 2020. — Vol. 14. — P. 241–258.

[26] Dietterich Thomas G et al. Ensemble learning // The handbook of brain theory and neural networks. — 2002. — Vol. 2, no. 1. — P. 110–125.

[27] Sagi Omer, Rokach Lior. Ensemble learning: A survey // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. — 2018. — Vol. 8, no. 4. — P. e1249.

[28] Zhou Zhi-Hua, Zhou Zhi-Hua. Ensemble learning. — Springer, 2021.

[29] Zhang G.Peter. Time series forecasting using a hybrid ARIMA and neural network model // Neurocomputing. — 2003. — Vol. 50. — P. 159–175.

[30] Palani Sundarambal, Liong Shie-Yui, Tkalich Pavel. An ANN application for water quality forecasting // Marine Pollution Bulletin. — 2008. — Vol. 56, no. 9. — P. 1586–1597.

[31] Sherstinsky Alex. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network // Physica D: Nonlinear Phenomena. — 2020. — Vol. 404. — P. 132306.

[32] Dey Rahul, Salem Fathi M. Gate-variants of Gated Recurrent Unit (GRU) neural networks // 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). — 2017. — P. 1597–1600.

[33] Manaswi Navin Kumar. RNN and LSTM // Deep Learning with Applications Using Python : Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras. — Berkeley, CA : Apress, 2018. — P. 115–126.

[34] De Gooijer Jan G., Hyndman Rob J. 25 years of time series forecasting // International Journal of Forecasting. — 2006. — Vol. 22, no. 3. — P. 443–473. — Twenty five years of forecasting.

[35] Deep Learning for Time Series Forecasting: A Survey / José F. Torres, Dalil Hadjout, Abderrazak Sebaa et al. // Big Data. — 2021. — Vol. 9, no. 1. — P. 3–21.

[36] Sezer Omer Berat, Gudelek Mehmet Ugur, Ozbayoglu Ahmet Murat. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019 // Applied Soft Computing. — 2020. — Vol. 90. — P. 106181.

[37] Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research // Journal of Pharmaceutical and Biomedical Analysis. — 2000. — Vol. 22, no. 5. — P. 717–727.

[38] Lim Bryan, Zohren Stefan. Time-series forecasting with deep learning: a survey // Philosophical Transactions of the Royal Society A. — 2021. — Vol. 379, no. 2194. — P. 20200209.

[39] An empirical comparison of machine learning models for time series forecasting / Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, Hisham El-Shishiny // Econometric reviews. — 2010. — Vol. 29, no. 5-6. — P. 594–621.

[40] Kolarik Thomas, Rudorfer Gottfried. Time series forecasting using neural networks // ACM Sigapl Apl Quote Quad. — 1994. — Vol. 25, no. 1. — P. 86–94.

[41] Yan Weizhong. Toward automatic time-series forecasting using neural networks // IEEE transactions on neural networks and learning systems. — 2012. — Vol. 23, no. 7. — P. 1028–1039.

[42] Sagi Omer, Rokach Lior. Approximating XGBoost with an interpretable decision tree // Information Sciences. — 2021. — Vol. 572. — P. 522–542.

[43] Experimenting XGBoost algorithm for prediction and classification of different datasets / Santhanam Ramraj, Nishant Uzir, R Sunil, Shatadeep Banerjee // International Journal of Control Theory and Applications. — 2016. — Vol. 9, no. 40. — P. 651–662.

[44] Chen Tianqi, Guestrin Carlos. Xgboost: A scalable tree boosting system // Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. — 2016. — P. 785–794.

[45] Xgboost: extreme gradient boosting / Tianqi Chen, Tong He, Michael Benesty et al. // R package version 0.4-2. — 2015. — Vol. 1, no. 4. — P. 1–4.

[46] Ensemble learning for data stream analysis: A survey / Bartosz Krawczyk, Leandro L Minku, João Gama et al. // Information Fusion. — 2017. — Vol. 37. — P. 132–156.

[47] Lightgbm: A highly efficient gradient boosting decision tree / Guolin Ke, Qi Meng, Thomas Finley et al. // Advances in neural information processing systems. — 2017. — Vol. 30.

[48] Al Daoud Essam. Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset // International Journal of Computer and Information Engineering. — 2019. — Vol. 13, no. 1. — P. 6–10.

[49] Hancock John T, Khoshgoftaar Taghi M. CatBoost for big data: an interdisciplinary review // Journal of big data. — 2020. — Vol. 7, no. 1. — P. 1–45.

[50] CatBoost: unbiased boosting with categorical features / Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev et al. // Advances in neural information processing systems. — 2018. — Vol. 31.

[51] Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions / Guomin Huang, Lifeng Wu, Xin Ma et al. // Journal of Hydrology. — 2019. — Vol. 574. — P. 1029–1041.

[52] A communication-efficient parallel algorithm for decision tree / Qi Meng, Guolin Ke, Taifeng Wang et al. // Advances in Neural Information Processing Systems. — 2016. — Vol. 29.

[53] Quickly boosting decision trees–pruning underachieving features early / Ron Appel, Thomas Fuchs, Piotr Dollár, Pietro Perona // International conference on machine learning / PMLR. — 2013. — P. 594–602.

[54] Pan Bingyue. Application of XGBoost algorithm in hourly PM2. 5 concentration prediction // IOP conference series: earth and environmental science / IOP publishing. — Vol. 113. — 2018. — P. 012127.

[55] A model combining convolutional neural network and LightGBM algorithm for ultra-short-term wind power forecasting / Yun Ju, Guangyu Sun, Quanhe Chen et al. // Ieee Access. — 2019. — Vol. 7. — P. 28309–28318.

[56] Dorogush Anna Veronika, Ershov Vasily, Gulin Andrey. CatBoost: gradient boosting with categorical features support // arXiv preprint arXiv:1810.11363. — 2018.

[57] Marcílio Wilson E, Eler Danilo M. From explanations to feature selection: assessing SHAP values as feature selection mechanism // 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI) / Ieee. — 2020. — P. 340–347.

[58] What makes an online review more helpful: an interpretation framework using XGBoost and SHAP values / Yuan Meng, Nianhua Yang, Zhilin Qian, Gaoyu Zhang // Journal of Theoretical and Applied Electronic Commerce Research. — 2020. — Vol. 16, no. 3. — P. 466–490.

[59] Mokhtari Karim El, Higdon Ben Peachey, Başar Ayşe. Interpreting financial time series with SHAP values // Proceedings of the 29th annual international conference on computer science and software engineering. — 2019. — P. 166–172.

[60] Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods / Dong Wang, Sven Thunéll, Ulrika Lindberg et al. // Journal of Environmental Management. — 2022. — Vol. 301. — P. 113941.

[61] Sundararajan Mukund, Najmi Amir. The many Shapley values for model explanation // International conference on machine learning / PMLR. — 2020. — P. 9269–9278.

[62] Ghorbani Amirata, Zou James. Data shapley: Equitable valuation of data for machine learning // International conference on machine learning / PMLR. — 2019. — P. 2242–2251.

[63] Rodríguez-Pérez Raquel, Bajorath Jürgen. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions // Journal of computer-aided molecular design. — 2020. — Vol. 34. — P. 1013–1026.

[64] Winter Eyal. The shapley value // Handbook of game theory with economic applications. — 2002. — Vol. 3. — P. 2025–2054.

[65] Roth Alvin E. The Shapley value: essays in honor of Lloyd S. Shapley. — Cambridge University Press, 1988.

[66] Monderer Dov, Samet Dov. Variations on the Shapley value // Handbook of game theory with economic applications. — 2002. — Vol. 3. — P. 2055–2076.

[67] Towards efficient data valuation based on the shapley value / Ruoxi Jia, David Dao, Boxin Wang et al. // The 22nd International Conference on Artificial Intelligence and Statistics / PMLR. — 2019. — P. 1167–1176.

[68] Algorithms to estimate Shapley value feature attributions / Hugh Chen, Ian C Covert, Scott M Lundberg, Su-In Lee // Nature Machine Intelligence. — 2023. — P. 1–12.

[69] Fryer Daniel, Strümke Inga, Nguyen Hien. Shapley values for feature selection: The good, the bad, and the axioms // Ieee Access. — 2021. — Vol. 9. — P. 144352–144360.

[70] Problems with Shapley-value-based explanations as feature importance measures / I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, Sorelle Friedler // International Conference on Machine Learning / PMLR. — 2020. — P. 5491–5500.

[71] Littlechild Stephen C, Owen Guillermo. A simple expression for the Shapley value in a special case // Management Science. — 1973. — Vol. 20, no. 3. — P. 370–372.

[72] Kalai Ehud, Samet Dov. On weighted Shapley values // International journal of game theory. — 1987. — Vol. 16. — P. 205–222.

[73] Hart Sergiu, Mas-Colell Andreu et al. The potential of the Shapley value // the Shapley value. — 1988. — P. 127–137.

[74] The shapley value in machine learning / Benedek Rozemberczki, Lauren Watson, Péter Bayer et al. // arXiv preprint arXiv:2202.05594. — 2022.

[75] Merrick Luke, Taly Ankur. The explanation game: Explaining machine learning models using shapley values // Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4 / Springer. — 2020. — P. 17–38.

[76] Mohseni Sina, Zarei Niloofar, Ragan Eric D. A multidisciplinary survey and framework for design and evaluation of explainable AI systems // ACM Transactions on Interactive Intelligent Systems (TiiS). — 2021. — Vol. 11, no. 3-4. — P. 1–45.

[77] Towards a rigorous evaluation of XAI methods on time series / Udo Schlegel, Hiba Arnout, Mennatallah El-Assady et al. // 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW) / IEEE. — 2019. — P. 4197–4201.

[78] Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model / Basim Mahbooba, Mohan Timilsina, Radhya Sahal, Martin Serrano // Complexity. — 2021. — Vol. 2021. — P. 1–11.

[79] Liao Q Vera, Gruen Daniel, Miller Sarah. Questioning the AI: informing design practices for explainable AI user experiences // Proceedings of the 2020 CHI conference on human factors in computing systems. — 2020. — P. 1–15.

[80] Gunning David, Aha David. DARPA's explainable artificial intelligence (XAI) program // AI magazine. — 2019. — Vol. 40, no. 2. — P. 44–58.

[81] Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI / Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser et al. // Information fusion. — 2020. — Vol. 58. — P. 82–115.

[82] Das Arun, Rad Paul. Opportunities and challenges in explainable artificial intelligence (xai): A survey // arXiv preprint arXiv:2006.11371. — 2020.

[83] Explainable artificial intelligence (XAI) in deep learning-based medical image analysis / Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, Max A Viergever // Medical Image Analysis. — 2022. — Vol. 79. — P. 102470.

[84] Adadi Amina, Berrada Mohammed. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) // IEEE access. — 2018. — Vol. 6. — P. 52138–52160.

[85] Tjoa Erico, Guan Cuntai. A survey on explainable artificial intelligence (xai): Toward medical xai // IEEE transactions on neural networks and learning systems. — 2020. — Vol. 32, no. 11. — P. 4793–4813.

[86] What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research / Markus Langer, Daniel Oster, Timo Speith et al. // Artificial Intelligence. — 2021. — Vol. 296. — P. 103473.

[87] Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review / Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz et al. // Applied Sciences. — 2021. — Vol. 11, no. 11. — P. 5088.

[88] Liao Q Vera, Varshney Kush R. Human-centered explainable ai (xai): From algorithms to user experiences // arXiv preprint arXiv:2110.10790. — 2021.

[89] Argumentative XAI: a survey / Kristijonas Čyras, Antonio Rago, Emanuele Albini et al. // arXiv preprint arXiv:2105.11266. — 2021.

[90] Saeed Waddah, Omlin Christian. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities // Knowledge-Based Systems. — 2023. — Vol. 263. — P. 110273.

[91] Evaluating XAI: A comparison of rule-based and example-based explanations / Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, Mark Neerincx // Artificial Intelligence. — 2021. — Vol. 291. — P. 103404.

[92] Wolf Christine T. Explainability scenarios: towards scenario-based XAI design // Proceedings of the 24th International Conference on Intelligent User Interfaces. — 2019. — P. 252–257.

[93] Páez Andrés. The pragmatic turn in explainable artificial intelligence (XAI) // Minds and Machines. — 2019. — Vol. 29, no. 3. — P. 441–459.

[94] Explainable artificial intelligence: a comprehensive review / Dang Minh, H Xiang Wang, Y Fen Li, Tan N Nguyen // Artificial Intelligence Review. — 2022. — P. 1–66.

[95] Explainable artificial intelligence (xai) on timeseries data: A survey / Thomas Rojat, Raphaël Puget, David Filliat et al. // arXiv preprint arXiv:2104.00950. — 2021.

[96] Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities / R Machlev, L Heistrene, M Perl et al. // Energy and AI. — 2022. — Vol. 9. — P. 100169.

[97] Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies / Eoin M Kenny, Courtney Ford, Molly Quinn, Mark T Keane // Artificial Intelligence. — 2021. — Vol. 294. — P. 103459.

[98] Dongfang Qi et al. Statistical analysis of investment attractiveness of China's regions // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2022. — Vol. 18, no. 1. — P. 188–194.

[99] Qi Dongfang, Bure Vladimir M. Research of investment attractiveness based on cluster analysis // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2023. — Vol. 19, no. 2. — P. 199–211.

[100] He Yang, Qi Dongfang, Bure Vladimir M. New application of multiple linear regression method-A case in China air quality // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2022. — Vol. 18, no. 4. — P. 515–526.

[101] He Yang, Qi Dongfang, Bure VM. Long-Term Air Quality Evaluation System Prediction In China Based On Multinomial Logistic Regression Method // GEOGRAPHY, ENVIRONMENT, SUSTAINABILITY. — 2024. — Vol. 16, no. 4. — P. 164–171.

[102] D. Qi V. M. Bure. Explanatory comparative analysis of time series forecasting algorithms for air quality prediction // Vestnik of Saint Petersburg University. Applied Mathematics. Computer Science. Control Process. — 2024. — Vol. 20, no. 2.

[103] ShapTime: A General XAI Approach for Explainable Time Series Forecasting / Yuyi Zhang, Qiushi Sun, Dongfang Qi et al. // Intelligent Systems Conference / Springer. — 2023. — P. 659–673.