

Enhancing the Forecasting Accuracy of Direct Tax using a Hybrid  
Approach and Machine Learning Models



By

Hiba Aftab

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

---

**Enhancing the Forecasting Accuracy of Direct Tax using a Hybrid  
Approach and Machine Learning Models**



By

**Hiba Aftab**

*A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR  
THE DEGREE OF MASTER OF PHILOSOPHY IN STATISTICS*

Supervised By

**Prof. Dr. Ijaz Hussain**

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

*In the Name of Allah The Most Merciful and The Most Beneficent*

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

## CERTIFICATE

Enhancing the forecasting accuracy of direct tax using a  
hybrid approach and machine learning models


By


Hiba Aftab


(Reg. No. 02222113009)

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE  
REQUIREMENTS FOR THE DEGREE OF M.PHIL. IN  
STATISTICS

*We accept this thesis as conforming to the required standards*

1.   
Prof. Dr. Ijaz Hussain  
(Supervisor)

2.   
Prof. Dr. Muhammad Hanif  
(External Examiner)

3.   
Prof. Dr. Ijaz Hussain  
(Chairman)

DEPARTMENT OF STATISTICS  
QUAID-I-AZAM UNIVERSITY  
ISLAMABAD, PAKISTAN  
2023

# Dedication

*I feel great honor and pleasure to dedicate this research work to*

***My Beloved Mother (Late) and My Father***

*whose endless support, prayers, and wishes have been my guiding light during my whole  
educational period.*

## Acknowledgment

First and foremost, I praise and acknowledge Allah Almighty, the Lord and creator of the heavens and earth. All respect and gratitude goes to the Holy Prophet Muhammad (Peace be upon him) who enlightens our hearts with the light of Islam and whose way of life has been a continuous guide for us deem it my utmost pleasure to avail this opportunity and express gratitude and a deep sense of obligation to my supervisor **Prof. Dr. Ijaz Hussain** for his valuable and dexterous guidance, scholarly criticism, untiring help, compassionate attitude, kind behavior, and moral support. I am greatly indebted to him for helping and guiding me throughout the thesis in every aspect.

I offer my deepest sense of gratitude, a profound respect and tribute to **Dr. Manzoor Khan, Dr. Abdul Haq, Dr. Sajid, Dr. Ismail, and Mam Maryam** for enlightening my statistical knowledge throughout my academic career and guidance throughout my research work. My acknowledgment would be incomplete without thanking the biggest source of my strength, my beloved parents, who have supported me financially as well as morally, and without their proper guidance, it would have been impossible for me to complete my higher education. I feel great honor to dedicate this research work to my **Nana Abu**, whose blessings and prayers have always been with me during my whole educational period. I wish to record my sincere appreciation to my seniors, classmates, and friends for their cooperation and help, especially **Swera Zeb Abbasi, Waleeja-tur-Rabbia, Amna Bibi, Saira Baig, Zoha Fahim, Hina Tariq, Faizan Bin Sabir, Ghulam Zainab, Syeda Aleen Zehra Bukhari, Mehnaz Bibi**. Last but not least, I feel pleasure to acknowledge those who love me and whom I love. Hopefully, this M.Phil dissertation will not be the end of my journey in seeking more knowledge to understand the meaning of life.

**Hiba Aftab**

## Abstract

Direct taxes have a significant impact on an economy, both in terms of income collection for the government as well as their consequences on individuals, enterprises, and overall economic behavior. Accurate forecasting of direct tax revenue is vital for effective fiscal planning and policy formulation. Traditional forecasting approaches frequently fail to reflect the complex patterns and financial datasets. We applied machine learning and hybrid methodologies to improve the prediction accuracy of direct tax revenue estimates. Univariate time series models, which include the Autoregressive Integrated Moving Average (ARIMA), Exponential Smoothing State Space (ETS), and Radial Basis Function Neural Network (RBFNN) model, have been used. Two-stage hybrid models WA-ARIMA, WA-ETS, WA-RBFNN, EMD-ARIMA, EMD-ETS, EMD-RBFNN, which are based on denoising and decomposition techniques, i.e., Wavelet Analysis (WA) and Empirical Mode Decomposition (EMD). Three-stage hybrid models including WA-CEEMDAN-ARIMA, WA-CEEMDAN-ETS, and WA-CEEMDAN-RBFNN are created based on the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) technique, and these models are applied using both WA and EMD. Comparison is made on the basis of Mean Square Error (MSE). Hence, in terms of prediction accuracy of direct tax, the WA-CEEMDAN-ARIMA has the lowest MSE value as compared to all other existing models. Hence, results show that WA-CEEMDAN-ARIMA is proved to be appropriate for this time series data.

# Contents

<b>1</b>	<b>Background of Study</b>	<b>1</b>
1.1	Introduction . . . . .	1
1.2	Research Objective . . . . .	9
1.3	Thesis at Glance . . . . .	9
<b>2</b>	<b>Univariate Linear Time Series Analysis</b>	<b>10</b>
2.1	Introduction . . . . .	10
2.2	Study Area . . . . .	17
2.3	Box-Jenkins Methodology . . . . .	17
2.4	Autoregressive Integrated Moving Average Model . . . . .	18
2.4.1	Autoregressive Model . . . . .	19
2.4.2	Integrated Process . . . . .	20
2.4.3	Moving Average Model . . . . .	21
2.4.4	Autoregressive Moving Average Model . . . . .	21
2.4.5	Autocorrelation and Partial Autocorrelation Functions . . . . .	22
2.4.6	Augmented Dickey-Fuller Test . . . . .	22
2.5	Exponential smoothing State Space Model . . . . .	23
2.6	Radial Basis Function Neural Network Model . . . . .	24
2.7	Result and Discussion . . . . .	25
2.8	Identifying Model parameters . . . . .	26
2.9	Conclusion . . . . .	30
<b>3</b>	<b>Decomposition Techniques and Hybrid Models</b>	<b>32</b>
3.1	Introduction . . . . .	32
3.1.1	Continuous Wavelet Transform . . . . .	35
3.1.2	Discrete Wavelet Transform . . . . .	35
3.2	Techniques of Decomposition . . . . .	36
3.2.1	Empirical Mode Decomposition . . . . .	37
3.2.2	Intrinsic Mode Functions . . . . .	37
3.2.3	Steps of Decomposition into Intrinsic Mode Functions . . . . .	38
3.3	Results and Discussion . . . . .	39
3.4	Conclusions . . . . .	48



---

<b>4</b>	<b>Three stage models</b>	<b>50</b>
4.1	Introduction . . . . .	50
4.2	Decomposition Techniques . . . . .	53
4.2.1	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise	53
4.3	Results and Discussion . . . . .	56
4.4	Conclusion . . . . .	63
<b>5</b>	<b>Conclusions and Recommendations</b>	<b>66</b>
5.1	Conclusion . . . . .	66
5.2	Recommendations . . . . .	68
	<b>Reference</b>	<b>70</b>

# List of Tables

2.1	Identification of AR and MA order through ACF and PACF . . . . .	22
2.2	ADF test for checking stationarity of direct tax data of time series. . . . .	25
2.3	RBFNN showing short-term forecasting of direct tax . . . . .	29
2.4	Comparison of mean square error of ARIMA, ETS, and RBFNN model . . . . .	30
3.1	WA-ARIMA model for the direct taxes . . . . .	40
3.2	WA-ETS model showing additive error, additive damped trend and error, trend parameters . . . . .	41
3.3	The evaluation of the prediction error of suggested WA-ARIMA, WA-ETS, WA-RBFNN in comparison with one-stage models . . . . .	41
3.4	The evaluation of the prediction error of suggested WA-ARIMA, WA-ETS, WA-RBFNN, EMD-ARIMA, EMD-ETS, EMD-RBFNN in comparison with one-stage models . . . . .	48
4.1	The evaluation of the prediction error of proposed models WA-CEEMDAN-ARIMA, WA-CEEMDAN-ETS, WA-CEEMDAN-RBFNN and EMD-CEEMDAN-ARIMA, EMD-CEEMDAN-ETS, EMD-CEEMDAN-RBFNN in comparison with all one-stage and two-stage hybrid model . . . . .	64

# List of Figures

2.1	Design of Radial Basis Function Neural Network . . . . .	25
2.2	Plot showing trend and seasonality in original series of direct tax revenues . . . . .	26
2.3	Correlogram of series for Direct Tax time series data . . . . .	27
2.4	Plot for best ETS model, showing observe value and slope along with trend and seasonality component: . . . . .	29
3.1	The Denoised series of the direct tax obtained by WA(in purple color) . . . . .	39
3.2	The EMD decomposition of the direct tax into six IMFs and one residual. . . . .	42
3.3	After applying the threshold, the smooth EMD decomposition of the direct tax into six IMFs and one residual. . . . .	43
3.4	The reconstructed series, the decomposed output obtained through EMD with threshold method (in blue color) . . . . .	44
3.5	The denoised series for direct tax. The denoised is obtained by wavelet (in purple color) and EMD (in green color) . . . . .	45
3.6	Prediction results of original series of direct tax in comparison with EMD-ARIMA (in red color) . . . . .	46
3.7	Prediction results of original series of direct tax in comparison with EMD-ETS (in green color) . . . . .	47
4.1	The EMD-CEEMDAN decomposition of the direct tax into five IMFs and one residual. . . . .	56
4.2	After applying the threshold, the smooth EMD-CEEMDAN decomposition of the direct tax into five IMFs and one residual. . . . .	58
4.3	The WA-CEEMDAN decomposition of the direct tax into six IMFs and one residual . . . . .	59
4.4	After applying the threshold, the smooth WA-CEEMDAN decomposition of the direct tax into five IMFs and one residual . . . . .	60
4.5	Prediction results of original series of direct tax in comparison with ARIMA (in blue color), EMD-ARIMA (in red color) and EMD-CEEMDAN-ARIMA (in green color) . . . . .	61
4.6	Prediction results of original series of direct tax in comparison with ETS (in blue color), EMD-ETS (in red color), and EMD-CEEMDAN-ETS (in green color) . . . . .	62

# Chapter 1

## Background of Study

### 1.1 Introduction

The word 'Revenue' originated from a French word meaning "to return." The word revenue was used in the perspective of money returning to financial resources, specifically in the form of taxes and other financial receipts in the ancient times of Rome. With the passage of time, the term 'revenue' evolved and expanded its scope mainly in the concepts of income as well as financial resources. The word 'Tax' comes from the Latin term 'taxare,' which means to analyze, estimate, or value. The background of taxation is quite complex and varies from time to time. Perhaps, Revenues and taxes perform a vital role in the functioning of any government and society. Revenues generated or produced through taxes are the preliminary source of funding at all levels for governments, whether it's local, regional, or national. Tax revenue collection is the process of collecting taxes from every single person, business, or other entity by the government or its authorized agencies. The tax revenue collection process basically involves the main steps, which are tax assessment, tax filing, tax payment, tax compliance monitoring, tax collection enforcement, record-keeping, and reporting. Tax revenues are used to manage essential public services, which include healthcare, education, public safety, transportation, and many other fields. There are different types of tax collection. i.e., Direct tax, sales tax, federal excise duty, and customs services. Direct tax is the major type in the field of tax revenues. The Federal Board of Revenue (FBR) plays a pivotal role in revenue collection for the government. However, the traditional methods and models employed for predicting direct tax revenue have often fallen short in providing precise forecasts, leading to

budgetary discrepancies and challenges in resource allocation. A fair and effective tax system is significantly dependent on the administration and collection of direct taxes in Pakistan. However, there are issues with efficiency and consistency with the current procedures and techniques used by the Federal Board of Revenue (FBR) for direct tax collection. These difficulties lead to postponed assessments, significant income loss, and increased taxpayer compliance burden. In addition, manual tax assessment procedures are subject to human mistakes, subjectivity, and inconsistencies. Many governments all around the world utilize direct taxes to collect revenue and fund public services. Time series analysis techniques have been applied by many nations but the application of forecasting techniques on direct taxes can be challenging due to various reasons such as the complex nature of a direct tax, volatility and uncertainty, behavioral response, policy changes, and legislative reforms, data limitations. Hence, there would be a need to check linearity and variation in the behavior of direct tax data. It is the major requirement to look for better forecasting techniques for modeling the tax revenue time series data. This research study seeks to contribute to tax revenue forecasting by investigating how a hybrid approach, combining traditional methods and machine learning models, can significantly enhance the accuracy of direct tax revenue forecasts at the Federal Board of Revenue. This study aims to provide actionable insights and recommendations for improving revenue projections with the help of taxes. Forecasting direct taxes is motivated by the need for effective fiscal planning and economic management. Governments can better plan for budgetary needs, debt management, and the funding of essential public services and infrastructure by forecasting future direct tax revenues. Accurate forecasts also support budgetary accountability and transparency, boosting trust among taxpayers and investors. The main objective of forecasting direct taxes is to give governments, politicians, and financial institutions an accurate and data-driven forecast of future tax revenues.

The first [Dietsch and Rixen \(2015\)](#) examined the impact of three different types of tax rivalries on the de facto sovereignty of states. It demonstrated how societal inequities were made worse by tax competitiveness to demonstrate why it was an instance of background injustice. Additionally, this research addressed the criticism that supporting a cosmopolitan view of global justice would be in conflict with these principles. Further, It was claimed that the two ideas act as a constitutional toolbox to outline the degree to which interdependence between

states in budgetary matters was necessary for constitutional interdependence. [Scott \(1955\)](#) examined tax laws and systems in the context of a market economy. Complex tax systems with features that follow a variable but predictable pattern are produced by democratic institutions. The authors employed formal voting behavior modeling to create this research, highlighting recent developments in the theory of probabilistic voting. This study explained the existing tax literature by explicitly connecting fiscal decisions to voting and by looking at a democratic nation's tax systems from a number of angles. While describing observed characteristics of tax systems was the authors' main goal, they also spent a lot of time discussing the welfare and efficiency impacts of taxes in the presence of collective choice, as well as reviewing alternative models and relevant research. In order to connect theory to empirical data, they also used computational general equilibrium analysis and statistical research on federal and state governments in the US and Canada. In recent years, direct taxes have played a significant role in generating revenues by governments worldwide and promoting equity, as well as addressing fiscal challenges. Direct tax, for example, income tax, corporate tax as well as property tax, contributed a significant portion of government revenue. Here, the income tax was the typical kind of direct taxation charged on the income of a person or an organization. Usually, it was imposed on several kinds of revenue sources, including but not limited to salaries and wages, business gains, rental revenue, capital profits from asset sales, Income from dividends, and interest ([Salanie, 2011](#)). Based on the taxpayer's source of income and applicable tax system, the income taxation rates and tax brackets might differ significantly. Higher-income groups were often subject to higher tax rates under progressive tax policies ([Saad, 2014](#)). For instance, depending on the income level, income tax rates in various nations might vary from 0% to over 50% ([Piketty, 2015](#)). Profits made by companies or businesses are subject to corporate tax. Corporate tax rates differ from country to country and may be affected by elements like the nature of the business, the size of the organization, and the industry. Generally, Corporate tax rates vary from 15% to 35%; however, they might be higher or lower depending on the nation's tax policies ([Atwood et al., 2012](#)). Now, the other kind of Direct Tax is capital gains tax. Revenue made from the sale or disposal of capital resources, such as real estate, stocks, bonds, or expensive personal property, liable to capital gains tax. Rates of capital gains taxation might differ

from income tax rates. They were often created to promote long-term investments. The rates might vary from zero percent (0%), which applied to certain low-income levels, to higher percentages 15 – 20%, for higher income levels (Alvaredo et al., 2013). Depending on the amount of assets or net worth that a person or family has, certain nations might apply a wealth tax. Real estate, financial assets, automobiles, jewelry, and other costly belongings may all be subject to the wealth tax, which has broad range of possible rates. Wealth taxes are levied in a number of nations, and the percentage rates may vary from extremely low to high percentages. now, consider the following examples of different countries to have a concept of the wealth tax rates (Zucman, 2015). In France, a real estate wealth tax took the role of wealth tax in 2018. A progressive wealth tax rate ranging from 0.5% to 1.5% on an individual's entire net worth was imposed under the former system on those whose net taxable wealth exceeded €1.3 million (Krenek and Schratzenstaller, 2018). In Norway, Individuals having net taxable wealth that exceeded a certain level were liable for a wealth tax. The progressive revenue collection rate depends on the taxpayer's net worth and varies from 0.15% to 0.7% (Banoun, 2020). Spain also imposed a wealth tax referred to as the "Impuesto sobre el Patrimonio." Since this revenue collection was governed at the regional level, the rates vary throughout the various regions of Spain. The rates on net worth that exceeded certain thresholds might vary from 0.2% to 3.5% (Kapeller et al., 2021). In Switzerland, the wealth tax rates vary widely amongst the cantons (regions). The rates might vary based on the canton and the taxpayer's net wealth from 0.01% to 1% or more (Martinez, 2017). In terms of GDP, direct tax income reached 4.3 percent in FY18, up from 2.9 percent on average from FY03 to FY17. Direct taxes' share of the FBR's income might shift from year to year depending on the economy, tax laws, and other variables. The cumulative direct tax payment to the FBR ceased in September 2021. In FY 2016-2017, a significant amount of the FBR's income came from direct taxes. Although the precise proportion may change, it normally represents a significant portion of the overall tax income received by the FBR. In FY 2017–2018, direct taxes kept being a major source of income for the FBR. They played a significant role in the government's efforts to finance public spending and achieve a number of development and social welfare goals. Direct taxes remained the FBR's basic source of income during FY 2018-2019. Improved compliance and a wider tax base were the goals of

the government's tax policy and enforcement actions, which increased the direct taxes' share of the FBR's total income. As in FY 2018-2019, in FY 2019-2020 the direct taxes continued to provide a significant contribution to the FBR's revenue. The government placed a number of efforts and changes to make the tax system simpler, increase transparency, and promote voluntary tax compliance. In FY 2020 – 2021, the collection of direct taxes was affected due to COVID-19. The actual contribution in this particular year may change depending on the state of the economy and the steps the government takes to solve the problems caused by the epidemic (FBR, 2023). To encourage economic cooperation, avoid double taxation, and communicate tax-related information, Pakistan signed tax treaties, also known as double taxation avoidance agreements (DTAAs), with several nations. These agreements guaranteed a fair and effective tax system while facilitating trade and investment (Ali et al., 2022). The following important nations had tax treaties with Pakistan. A tax agreement exists between Pakistan and the United States to avoid double taxation and advance business ties. The agreement addressed a number of topics, such as corporate earnings, dividends, interest, royalties, and capital gains. It also had conditions that allowed the two nations to share data on taxes. A tax agreement between Pakistan and the United Kingdom also resolved difficulties with double taxation. Business earnings, dividends, interest, royalties, and capital gains were all covered under the agreement. Additionally, it made information sharing between the tax authorities of the two nations easier. A tax agreement between Pakistan and China was established to eliminate double taxation and fiscal fraud. Business earnings, interest, royalties, and capital profits were several types of money that were covered by the treaty. Additionally, it offered procedures for addressing any conflicts that could develop between the two nations. A tax agreement between Pakistan and the UAE aims to reduce double taxation and enhance economic cooperation. Business earnings, dividends, interest, royalties, and capital gains were all covered under the agreement. Additionally, it made it easier for the two nations to share the details about taxes. Similarly, like other countries, Pakistan had signed an agreement with Saudi Arabia to prevent double taxation and avoid financial fraud. Business earnings, dividends, interest, royalties, and capital gains are all covered under this agreement. It also provided mechanisms for information cooperation between the two countries' tax departments. In recent years, Pakistan has signed tax treaties with other



nations, including Germany, France, Turkey, Malaysia, and many more. These agreements were essential for promoting international trade, generating foreign capital, and assuring a fair and open tax system for people and companies doing cross-border business. Foreign direct investment played an essential role in financial development, particularly in economies that were developing. It enhanced learning, medical care, and business, as well as produced more employment opportunities. Percentage of FDI inflows in Pakistan evolved every year. Several countries tried to reconsider their tax laws so as to lure more FDI by offering numerous tax benefits like refunds of taxes, capital allowances, deductions, and so on. [Shafiq et al. \(2021\)](#) investigated the function of taxation in choice-making for FDI transfers into Pakistan. Time series data used from 1985 to 2020. Information came through two ways: WDI and the PES. For empirical analysis, ARDL and ECM methodologies were utilized. Reduced revenues, based upon research, stimulate foreign investors to make a contribution towards investment, while there was a long-run association between taxes and investment in Pakistan. Further controlling factors included a rise in GDP, openness to trade, and the exchange value, a favorable influence on investments. It was advised by policymakers to target policies to cut taxes in order which invite investments into Pakistan. The administration should reevaluate objectives when enacting FDI-friendly legislation. [Azam and Lukman \(2010\)](#) evaluated the consequences for different economic aspects of foreign direct investment into Pakistan, India, and Indonesia from 1971 to 2005. The log-linear regression model was utilized for every nation, and the least squares method was used in evaluating influence of important financial factors regarding investment. Most significant economic factors of FDI, based on empirical data, are the size of markets, external debt, local expenditures, trade openness, and real estate are all important factors. Furthermore, research discovered the theoretical findings of India's financial factors mirrored the experimental results of Pakistan except for two variables; however, the findings of Indonesia are inconsistent with the conclusions of Pakistan and India's financial drivers of FDI. To attract greater foreign direct investment into Pakistan, India, and Indonesia, Administrative authorities must ensure political and financial stability, infrastructure service, prosperity safety, regulation of local investment stimulation to reduce foreign debt, as well as give similar weight to appropriate monetary and fiscal policy. [Munir and Sultan \(2018\)](#) examined taxation's impact on business growth, both the extended and

brief run. Analysis used a basic time series model with GDP as the centered element also various types of revenues as an independent element from 1976 to 2014 on a yearly basis for Pakistan. In the future, direct revenues have a favorable relationship with economic growth. Sales revenues, tariffs on foreign trade, and other indirect revenues have a favorable long-term and short-term impact on Pakistan's economic growth. However, after one year, sales tax as well as other indirect revenues, have a negative influence on revenue development because individuals see a decrease in their real salary. The administration might raise tax revenues by broadening the foundation of revenue. Additional revenues typically had an adverse aftereffect for one to two years, so the administration must reduce it was dependent on additional revenues. The authorities must encourage public understanding about taxes to increase taxation enthusiasm and broaden the revenue roots into direct and indirect revenues, with secondary revenues more divided into branches of revenues. This research could help policymakers design taxes and understand how they affect growth. The way that the private income tax was organized and social insurance transactions were modified for the number of families subject to study. It was focused on the difficulties in the analysis of the formulation of legislation increase as a result of limits on the instruments that can be utilized, disparities in public assessments regarding the merits of transfer, also disagreements regarding the disincentive effects of taxation. [Atkinson and Bourguignon \(1990\)](#) analyzed these concerns regarding the French system of tax benefits, as well as ramifications of the System in France being "harmonized" with various aspects of the British approach. In developing countries, the informal sector was larger than in wealthier countries. This was due to underdeveloped countries' greater fixed costs of entering the formal economy. According to [Auriol and Warlters \(2005\)](#), boosting entry obstacles was in line with a purposeful administrative objective to raise taxation fees. Commodities entrance income promoted the growth of significant contributors, producing commodities strength and thus leases approved entrants. Prices could be simply seized by the state at low administrative prices through entry fees and profits taxes. A 64-state sample was used to examine the theory's relevance. The findings of this study were supported by empirical evidence. [Hills et al. \(2016\)](#) studied spending patterns, equitable effects, as well as alterations in public security and direct tax programs impacts since 2007. It emphasized the significance of protecting the early post-crisis worth of

benefits and tax credits in actual money, as well as disparity in the handling of retirees vs. employed perks. Actual expenditures for pension benefits increased during the Partnership and Labour administrations it began declining during the Coalition. Substantial rises in income-free individual exemptions for taxation of income reasons in addition to focused reduction regressive in employed wages. These repercussions would be exacerbated by the actions of the incoming Conservative government. The consequences of the Coalition's significant Universal Credit also 'retirees freedom' measures are still unknown. Pakistan's economy has had brief spurts of economic expansion; however, this progress has not been sustained. Fiscal inefficiency is frequently highlighted as one of the causes of Pakistan's erratic growth. [Qasim et al. \(2015\)](#) examined the influence of fiscal consolidation on growth in this study. The connection between financial policy components and growth using annual data from 1976 to 2014 was evaluated. The conclusion was that the fiscal imbalance does not relate linearly to productivity. Furthermore, interest payments had a negative link with growth; thus, critical to reduce each finance charge and the initial deficit. Because the existing revenue laws never promoted productivity, the tax laws must be reformed so that they both promote growth as well while maintaining the economy. To stimulate growth, development spending should be increased while existing spending is reduced. [Kemal et al. \(2017\)](#) determined financial integration had a favorable aftereffect on economic rise employing a non-linear framework in Pakistan. Using data from 1976 to 2015, a link between the budget deficit and technological development was examined so the ideal level of budget deficit for promoting growth was computed. Outcomes demonstrated that, at the present level, financial deficit had a positive association with economic growth but an extremely higher fiscal deficit must be detrimental to raise. Given the unpredictable connection between the budget deficit and economic growth, Pakistan should practice fiscal discipline and keep its deficit under control. [Bilquees \(2004\)](#) studied the suppleness and stability of the taxation method from 1974-75 to 2003-04. Net revenue's volatility in relation not related to farming GDP was not connected to the entire GDP. In all, sales tax dominated in terms of revenue growth. Withholding tax, an indirect tax that was included in income tax is a significant factor. The coefficient decreases when the withholding tax is eliminated. Sales tax on imports and manufacturing made up for lost money as a result of tariffs and decreased taxation on excise. It had major adverse effects on

the poor including the service industry and utilities in the sales tax net, which was represented by the sales tax coefficient relative to the GDP base. Despite optimistic predictions, tax modifications did not significantly enhance revenue. The weak stability of income tax in which withholding taxes were not included demonstrated that increasing the taxable income limit along with imposing astronomical withholding taxes were at odds with one another.

## 1.2 Research Objective

The objectives of study are:

1. To evaluate the performance of traditional time series models, including ARIMA and ETS, in comparison to the machine learning model, the Radial Basis Function Neural Network (RBFNN) to improve the prediction accuracy of direct tax revenues.
2. To examine the application of Wavelet Analysis and Empirical Mode Decomposition (EMD) as preprocessing techniques that improve the prediction accuracy of direct tax revenue.
3. To explore the implementation of the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) technique in combination with Wavelet Analysis and EMD to achieve the most accurate results.

## 1.3 Thesis at Glance

This present study is divided into 5 chapters. Chapter 1 is the description of direct tax, its primary types, and its uses. Pakistan's tax agreements with other countries and direct tax control for the progress of Pakistan's economy are also described. Univariate time series models and their comparison are discussed in Chapter 2.

Chapter 3 explains all the denoising and decomposition techniques with the two stages, and Chapter 4 explains three-stage hybrid models, and a comparison between one, two, and three stages is also discussed. Overall results, as well as suggestions, are discussed in Chapter 5.

# Chapter 2

## Univariate Linear Time Series Analysis

### 2.1 Introduction

In this section, the univariate linear time series analysis is discussed. In the field of statistics and econometrics, it is a standard time series forecasting model used to examine as well as predict data points across time. In order to make the data stationary, ARIMA combines Autoregressive (AR) and Moving Average (MA) components with differencing. It is frequently used for forecasting in a variety of sectors, including environmental science, finance, and economics. ARIMA remains a valuable and commonly used tool in the field of time series forecasting due to its simplicity, ease of interpretation, and efficiency in modeling a variety of data patterns. It also serves as a conceptual basis for enhanced forecasting methods. Indeed, time series data are analyzed and forecasted using the ETS model by decomposing it into three essential elements: error, trend, and seasonality. This decomposition makes it possible to better analyze and model the underlying patterns in the data when working with time series data that exhibit systematic (trend and seasonality) and random (error) patterns across time. The ETS model aids in producing predictions and projections that are based on historical data that are more precise by recognizing and modeling these components. Radial Basis Function Neural Network (RBFNN) is perfect for a variety of applications due to its flexibility in handling both univariate and multivariate modeling and forecasting problems. The motivation to use RBFNN lies in their capacity to model complex, non-linearity in data. [Streimikiene et al. \(2018\)](#) estimated Pakistan's tax income for economic year 2016-2017 applying distinct time series methodologies as well as analyzing the effect of direct and indirect

taxes upon the working class. To evaluate how effective this research was, three distinct time series types as Autoregressive model, the Autoregressive Integrated Moving Average model (ARIMA), and the Vector Autoregression (VAR) model, were applied. Research showed the significance of evaluation also forecasting revenues to enhance the economic and fiscal policies. For the purpose of estimating overall tax collections, tax components were used. Conclusions showed that the ARIMA model provided the best forecasting values among these models for the total tax collection in Pakistan. Moreover, [Alfaki and Masih \(2015\)](#) studied Libya's oil refining data for the Naphtha product through forecasting models. The ARIMA method and Box Jenkins were used for forecasting stationary and time series that are not stationary. ARIMA(1,1,1) model was a highly suitable and effective one to use to describe time series data that meets the least values for AIC, SIC, and Mean Square Error (MSE) as well as the Boxljung test. The results provided the future image of actual monthly sales of Naphtha products. [Farhath et al. \(2016\)](#) proposed many vital models for enhancing the precision and effectiveness of time collection forecasting also simulations. The main intention was to provide brief summary of certain time series forecasting models that are well recognized and used in an application, as well as an overview of primary characteristics of utmost attention, the accuracy as well as efficiency in applying a model onto a dataset, the model was chosen to investigate prediction also assess the precision of various time series model layouts. Hence, the forecasting method ARIMA and Exponential smoothing techniques were used to forecast the best results. ARIMA was preferable to the exponential smoothing approach when the information was adequate and lengthy. Also, the correlation among previous assessments was consistent. [Hussain et al. \(2014\)](#) analyzed and created time series prediction models for Pakistan's direct and indirect tax income during an approximate 40-year period (1973–2011). Estimate of the share price has been a major subject in academics' interest over the years in the fields of finance and economics, who have an occupation to develop better predictive models. For time series prediction, autoregressive integrated moving average (ARIMA) models have been researched in the literature. [Ariyo et al. \(2014\)](#) examined the time-consuming process of developing an ARIMA-based stock price forecasting model. The Nigeria Stock Exchange, as well as the New York Stock Exchange, were used to apply the developed share exchange forecasting technique. The outcomes demonstrated that the ARIMA model

can successfully compete with current stock price predictions and has a high potential for short-term prediction. [Pack \(1990\)](#) examined a specific study from 1986 that appeared in the *International Journal of Forecasting* that univariate ARIMA time series modeling (Box-Jenkins) was not shown to be an enhanced accurate univariate time series forecasting technique unlike older and standard straightforward options, such as various exponential smoothing approaches, in light of several empirical investigations reported up within the estimating studies in the 1970s or 1980s. The study's result was that exponential smoothing, an improved time series was often approach than Box-Jenkins in forecasting department store sales, unfounded due to extremely small sample size. [Wang et al. \(2018\)](#) introduced a unique hybrid technique for estimating short-term ridership by using the wavelet decomposition and the autoregressive integrated moving average (ARIMA) model, two powerful signal processing tools. The wavelet decomposition was intended to portray the stochastic or occasionally radically altering elements of ridership patterns, whereas the seasonal ARIMA was used to depict rather constant and regular ridership patterns. The hybrid model had a significant advantage in spotting sudden changes in ridership patterns associated with certain train stations since it incorporates wavelet decomposition and reconstruction. The experimental results demonstrated that the hybrid method beats the individual ARIMA model for all types of ridership patterns, but it was particularly helpful for forecasting ridership at stations that commonly experience abrupt pattern changes as a result of special events. [Ramos et al. \(2015\)](#) evaluated how successfully ARIMA models and state space models predict the future. A case study of retail sales of five different types of women's footwear—boots, booties, flats, sandals, and shoes was used to demonstrate forecasting effectiveness. For both processes, the framework using the precise information from Akaike's Criteria in-sample period score was chosen from all viable models for additional examination in the out-of-sample period. Both single-step and multi-step forecasts were created. Results showed so, by utilizing an automated technique, the overall out-of-sample forecasting performance of state space and ARIMA models on one-step and multi-step forecasts, as evaluated by RMSE, MAE, and MAPE, was relatively equal. [Konarasinghe \(2019\)](#) concentrated on testing the SCM and ARIMA on forecasting the S&P BSE Sensex index and comparing their forecasting abilities. The Bombay Stock Exchange (BSE), one of the world's biggest share exchanges, was the first

Indian stock exchange. One of the primary research goals was the S&P BSE Sensex index, which evaluated the performance of the exchange's top 30 listed companies; fundamental analysis, technical analysis, and an artificial neural network were used to forecast the BSE Sensex. As a result, the Auto Regressive Integrated Moving Average (ARIMA) technique was found as a highly successful strategy for the goal among the examined models. The Sama Circular Model (SCM) is a technique for univariate forecasting. The SCM is capable of addressing these limitations. Different ARIMA models were tested to get the best forecast models for direct income tax and indirect federal revenue observations. The outcomes showed that ARIMA(1,1,1) and ARIMA(1,1,0) models were considered finest models with regard to direct income tax and indirect federal revenue accordingly. However, the Bayesian Vector Auto-regression (BVAR), Auto-regressive Moving Average (ARIMA), and State Space Exponential Smoothing (Error, Trend, Seasonal [ETS]) models used to model and forecast the major tax revenues of South Africa, including Corporate Income Tax (CIT), Personal Income Tax (PIT), Value-Added Tax (VAT), and Total Tax Revenue (TTR). Predictions of these three models relied on the Root mean square error (RMSE). The ETS showed the best results for TTR forecasting as it outperformed the BVAR method, (Molapo et al., 2019). To enhance a single model's forecasting capabilities in predicting inpatient admissions and hospital department income, Andellini et al. (2021) presented a hybrid ETS-ARIMA. The Bambino Gesù Children's Hospital (Italy) database on the General Surgery department was accessed between January 2012 and December 2018 to obtain statistics on monthly inpatient admissions and monthly revenues. The preliminary study's conclusions were encouraging: Hybrid models can outperform simple techniques with smaller errors. Hybrid models' relative errors were fewer than 5%, showing high forecasting ability. Finally, this research contributed to forecasting in the healthcare industry utilizing time series data. They might be applied to optimizing the hospital's decision-making procedures for scheduling inpatient admissions and allocating assets. Panigrahi and Behera (2017) created a hybrid forecasting approach that utilized a univariate forecasting model. ETS, a linear model, is combined with ANN, a non-linear model. By combining linear and nonlinear modeling skills in two methods, the hybrid model provides a more robust framework for discovering novel combinations of linear and nonlinear relationships. By ranking first, the outcomes demonstrated the dominance



of the suggested hybrid method. A meaningful input variable selection for the time series forecasting technique may be applied in place of the AR method for the lag selection of the ANN model, leading to improved model performance and for the deep levels to obtain the very best. In order to forecast nonseasonal time series data, [Wang et al. \(2013\)](#) suggested a hybrid model that integrates the time series ARIMA model with the neural network model. Together, the nonlinear ANNs model and the linear ARIMA model were utilized to examine time series data and pinpoint distinct trends. The hybrid model integrated ARIMA and ANNs' linear and nonlinear modeling skills. Results showed that the hybrid model outperformed the ARIMA, ANN, and additive models, which depended on analytical skills using three different datasets. For the purpose of evaluating the criterion, the accuracy metrics MSE, MAD, and MAPE were used. According to research findings, the hybrid model got the most accurate ratings.

[Benvenuto et al. \(2020\)](#) predicted the rapid spread of *COVID* – 19. ARIMA was used to forecast the physiological effects and occurrence of *COVID* – 19. The results indicated that the virus's growth rate is falling. For the improvement of better outcomes, case explanations and collection of data had to be obtained in person.

[Guha and Bandyopadhyay \(2016\)](#) forecasted the price of gold using the time series ARIMA model. ARIMA was used for forecasting upcoming prices of gold. ARIMA (1,1,1) is selected among models which fulfill all conditions of fit statistics. The limitations were that this model predicts only short-term changes in data. This model could only predict very short-term shifts in data. It got more challenging to identify a clear shift. It gets more challenging to identify a clear shift. This approach did not provide evidence that the gold price was linear, but this shortcoming might be worked around by using an ARIMA model, which satisfies the fundamental premise that forecasts should adhere to a linear trend. [Qurban et al. \(2021\)](#) developed two ground-breaking hybrid techniques meant to address the complex problems related to forecasting natural assets. These difficulties come from the frequently intricate interactions of variables, including multi-scale variability, nonlinearity, non-stationarity, and high irregularity, which were frequently seen in time series data from natural assets. A mixture of denoising, decomposition, prediction, and ensemble techniques served as the basis for the first of these unique hybrid approaches. This study's objective was to enhance the quality

of time-series data obtained from natural resources by expertly integrating these concepts. The precision of this first stage greatly enhanced the success of the following predictions. This involved looking at both conventional one-stage models with the missing denoising and decomposition processes and two-stage hybrid models with these techniques. They also delved into examining three-stage hybrid models, which used even more complex methods. The results showed the proposed framework for estimating natural assets. This convincing result showed that the newly proposed methodologies had to improve forecasting accuracy in the area of natural assets. In conclusion, a significant advancement mark in the field of forecasting for natural resources (Moh'd et al., 1998). A novel hybrid approach was proposed in a recent study headed by Panigrahi et al. (2021) for the goal of forecasting monthly and yearly sunspot time series (SN). The Auto-Regressive Integrated Moving Average (ARIMA), the Exponential Smoothing with Error, Trend, and Seasonality (ETS), and the Support Vector Machine (SVM) were merged in this method. The approach began by applying linear kernel functions-based ARIMA, ETS, and SVM models on the SN time series. The goal was to get the most projections possible from these models, which gave rise which called linear component forecasts. This stage mainly concentrates on identifying the more obvious patterns in the data. SVM deals with this nonlinearity, Gaussian kernel function is used. By selecting this kernel function, the residual series' complex, nonlinear interactions were handled with a sophisticated tool. The final forecasts were more complex than the simple sum of these factors. Finally, integrated the predictions from the linear and nonlinear components to get a final forecast. In conclusion, a study developed a complex hybrid forecasting approach that capitalized on the advantages of many modeling approaches. Idea and all study proved the quality of the method, demonstrating its potential to succeed across many vistas and accuracy measurements.

Gao et al. (2022) developed two unique approaches, Radial Basis Function Neural Network (RBFNN)-based and meta-heuristic computing paradigms for predicting PSS (Peak Shear Strength). The search for the RBFNN model's ideal parameters, which are essential for making correct predictions, was where the trip started. Ant Colony Optimization (ACO) and Grey Wolf Optimization (GWO) were two cleverly combined optimization techniques. These methods effectively functioned as scouts in the enormous parameter space, looking for

the ideal combination to maximize prediction potential. In the end, according to the data, the RBFNN-GWO model performed best, as evidenced by its remarkable R-squared (R<sup>2</sup>) value of 0.997. In addition to outstanding accuracy, it also displayed a noteworthy quality, i.e., faster convergence. The RBFNN-ACO and GEP models, in contrast, maintained their position with good R<sup>2</sup> values of 0.995 and 0.996, respectively. The RBFNN-GWO model served as a useful forecasting tool. Its ability to combine precision and efficiency made it a valuable instrument in the toolbox of experts engaged in the difficult profession of rock engineering. A fascinating look into the field of PSS prediction, where cutting-edge modeling approaches and optimization strategies came together to improve forecasting accuracy and simplify operational planning for rock engineers. Predictive modeling spanning six different subjects was the main focus of this study. [López-Martín \(2015\)](#) main goal was to evaluate the Radial Basis Function Neural Network's (RBFNN) capacity to forecast software project effort. The standard Regression Statistical Model (SLR), Feedforward Multilayer Perceptrons (MLP), which were frequently used to predict software project efforts, and General Regression Neural Networks (GRNN), a variant of RBFNN, were compared to RBFNN in order to accomplish this. The essential thing was whether the prediction accuracy of RBFNN would statistically outperform that of SLR, MLP, and GRNN when using adjusted function point software project data serving as the independent variable. During the training process, a cross-validation approach was used to evaluate these models and guarantee reliable findings. The absolute residuals were examined, and Friedman model-based statistical tests were performed. This study's surprising results showed a considerable difference in the four models' predicted accuracy, but there was a catch. With a 95% confidence level, the RBFNN outperformed SLR for brand-new software projects. [Jain and Mallick \(2017\)](#) on adjusting several elements to improve forecasting estimates. They understood models like Exponential Smoothing State Space (ETS) and AutoRegressive Integrated Moving Average (ARIMA) could be used to predict variables like air temperature, rainfall, and relative humidity. These models were tested, and it was discovered that they could accurately anticipate outcomes in a variety of situations and sufficiently explain the observed results. [Sun et al. \(2019\)](#) determined how long a train system will function, and a hybrid technique was created. A multi-layer decomposition method, where a Radial Basis Function Neural Network (RBFNN) played a crucial part in pre-

diction, was the key component of this approach. Due to the nonlinearity and nonstationarity that typically characterize deterioration series data, this method recognized the necessity for pretreatment. They used Improved Variational Mode Decomposition (IVMD) and Complete Ensemble Empirical Mode Decomposition (CEEMD) as preprocessing methods to deal with these problems. With the help of CEEMD, high-frequency intrinsic mode functions (IMFs) were obtained, also prediction accuracy was increased by further deconstructing the IMFs using IVMD. Then, RBFNN was used for all types of prediction problems.

## 2.2 Study Area

The study's primary asset is Direct tax collection in Pakistan. The four variables of tax revenues are Direct Tax(a tax that is applied directly on individuals or entities), Sales Tax(The sale of products and services are subject to a kind of consumption tax), Federal Excise Duty(the federal government's tax on a product, service, or industry inside the nation) and Custom Services(Services performed by the government or other authorized entities at international borders), the data of Direct Tax for the study is chosen. The data is time series data for a variable starting from the fiscal year 2003 and ending at the June of 2021. The main source of data is the Federal Board of Revenue of Pakistan (FBR)

## 2.3 Box-Jenkins Methodology

Box-Jenkins is an iterative technique that creates an ARIMA model for the seasonal and trend components, measures precise weighting parameters, tests the model, and repeats the process precisely. The Box-Jenkins technique was chosen to create simulation and forecasting systems that rely on The method's ability to handle complex situations, adaptability when correcting dependent time series, beneficial statistical and mathematical processes, effectiveness when dealing with risks due to programmability, and most importantly, the ease of use of the method. Any data set can use the forecasting model that Box-Jenkins offers. Additionally, it offers a structured way for developing, analyzing, and forecasting time series models as

part of its methodology. With this process, current observation serves as a starting point, and forecasting error is estimated for future projection. Only stationary time series could be used with the Box-Jenkins approach. A stagnant series is one that has neither seasonality nor a pattern of trends. The data used in this study is frequently non-stationary, so in order to make them stationary, first applied several transformation techniques [Lu and AbouRizk \(2009\)](#). The box-Jenkins approach suggests applying logarithmic or power transformations to achieve stationarity in variance and short and long (seasonal) to achieve stationarity in mean. According to [Nelson and Plosser \(1982\)](#), some series produced decent results with differencing, whereas others produced superior results with linear detrending. Most of the time, the series employed was non-stationary, making it challenging to apply any non-stationary series. Box-Jenkins ARIMA modeling provided a popular method for taking a non-stationary series' differences and turning them into a stationary series. AR, MA, or ARMA models could then easily fit the series. If stationarity in the mean was needed and the series exhibited seasonality, Box-Jenkins suggested seasonal models with long-term (seasonal) differencing.

## 2.4 Autoregressive Integrated Moving Average Model

Generally, a forecasting problem is resolved using linear methods such as ARMA and ARIMA models simplified by Box Jenkins (1967). General notation of model is:

$$Z_t = a_o + \sum a_i x_{t-i} + \sum b_j \epsilon_{t-j} \quad (2.1)$$

where  $i = 1, 2, \dots, p$  and  $j = 1, \dots, q$ , ([Tektaş, 2010](#)). Before working on the data, ARIMA model converted nonstationary data into stationary data by using Augmented Dickey-Fuller (ADF), Kwiatkowski-Phillips-Schmidt-Shin(KPSS), Phillips and Perron(PP) tests. These tests are used to check the stationary of the data. ARIMA technique was first developed by Box and Jenkins, which is why usually referred to as Box and Jenkins models. It has been utilized to do forecasting for linear data. The ARIMA technique provided significant flexibility in the recognition, parameter estimation, and forecasting of univariate time series models ([Mondal et al., 2014](#)). It has four stages in modeling: Identification, Estimation, Diagnostics, and Forecast.

The AR model shows a weighted moving average over prior observations, whereas the ARIMA technique relies on historical data. Similarly, Integrated(I) demonstrated a monotonic trend as well as a polynomial trend, and MA demonstrated a weighted moving average over past errors. Hence, ARIMA (p,d,q) was made up of three model parameters AR(p), I(d), and MA(q), where  $p$  is the order of autocorrelation,  $d$  is an order of differencing, and  $q$  is an order of MA. (Guha and Bandyopadhyay, 2016). This model included three models, i.e., AR, MA, and SARIMA. The parameters of ARIMA models were predicted by applying the ACF graph and PACF correlogram. The ARIMA and ETS models were used to make short-term predictions. The ETS was also used as a comparative forecasting method (Benvenuto et al., 2020) .

This model described the steps necessary to forecast the final cost for S & P500 index using a variety of statistical techniques, for example, the Autoregressive Integrated Moving Average (ARIMA) and the Exponential Smoothing (ETS) technique. In order to evaluate the models' precision, simplest strategy would be considered. Two techniques used in current research will be compared on the basis of standard deviation. The forecasting outcomes showed that the ARIMA model had a more accurate fit to the data, and might provide a positive general trend prediction when opposed to current approaches (Sun, 2020).

### 2.4.1 Autoregressive Model

An AR (Autoregressive) model is a type of time series model applied for analyzing and forecasting data that is time-dependent. In time series analysis, data is viewed as a series of observations, each of which is collected at a given time period. The AR model depends on the idea that the current index of a series  $x_t$  may be described by the prior values  $x_{t-1}, x_{t-2}, \dots, x_{t-p}$ , where  $p$  determines the number of prior steps required to anticipate the present value (Shumway et al., 2000). The nonstationary model, sometimes known as AR(1), is an essential approach in econometrics time series.

This method is represented by:

$$y_i = y_{i-1} + \mu_i \quad (2.2)$$

The above equation is shown as non drifted stationary model. If a constant value is inserted into the above equation, it becomes a nonstationary model with drift (Ghauri et al., 2020). In an autoregressive model, the time series values observed at central moments are the entire sum of the previous ones:

$$x_i = \mu + \sum \phi_i x_{t-i} + \mu_t x_i = \mu + \sum \phi_i x_{t-i} + \mu t \quad (2.3)$$

where  $t = 1, 2, 3, \dots, n$ .  $\mu_t$  is constant and is residual specifies a white noise process and  $p$  shows the order of the AR process. White noise follows the cross-sectional regression rules: the residual mean is zero, variance is constant, also there is no correlation between the error and white noise and also within variable  $x_{t-i}$  and the error  $\mu_t$ . In short, the expected value of  $x$  is  $\mu$ , as  $x_t$  is described as a weak stationary process. The variance of the series is fixed; moreover, it never relies upon time, which indicates the series is homoskedastic and the correlation relies just on the lag. The development of the regressive equation in which the prior one determined the value of  $t$ , is predicted for forecast because the upcoming value at the moment  $t + 1$  is clarified by the previous  $p$  time series value. It is simple to utilize original data  $x_i$  and back-transformed values  $\hat{x}_i$  to calculate the error and standard deviation. The estimated value gained apart from the model is also by the final  $p$  data known as pivotal or crucial values. The forecasted value  $\hat{x}_{n+1}$  is commonly represented by  $x_h$ , where  $h$  is known as the forecast horizon.

The next expected value determined using the predicted forecast value  $\hat{x}_h$ :

$$\hat{x}_{(2h)} = \phi_1 \hat{x}(h) + \phi_2 x_n + \dots + \phi_p x_{n-(p-1)} \quad (2.4)$$

by this procedure,  $m$  expected values  $\hat{x}_m h$  could be estimated successively (Werner, 2012).

## 2.4.2 Integrated Process

An order 1 integrated process is expressed as follows:

$$Y_t = Y(t - 1) + \epsilon_t \quad (2.5)$$

Where  $t$  is a white noise process, difference order 1 shows that there is no difference between two successive values of  $Y$ .

### 2.4.3 Moving Average Model

The concept of MA and AR was proposed by the work of Yule, Slutsky, Walker, and Yaglom (Mondal et al., 2014). Let that  $Y_t$  completely stochastic procedure with a zero mean as well as variance  $\sigma^2_z$ . Then procedure  $x_t$  is called as MA process of order  $n$  if

$$X_t = \beta_0 Y_t + \beta_1 Y_{t-1} + \dots + \beta_n Y_{t-n} \quad (2.6)$$

where  $\beta_i$  are constants. The  $Y$ 's are often scaled so that  $\beta_0=1$  (Chatfield et al., 2019). Typically, for the first order moving average method is:

$$X_t = Y_t + \beta Z_{t-1} \quad (2.7)$$

In time series, fitting an MA model is much fitting in comparison to generating an AR model because one of error random terminology in MA is unpredictable. Regardless of MA parameter values, the MA process remains stationary. If  $p=0$ , the procedure is known as MA (Moving Average) (Adhikari and Agrawal, 2013).

### 2.4.4 Autoregressive Moving Average Model

The ARMA model is a popular tool for analyzing time series (Mondal et al., 2014). An ARMA(p,q) model is a union of AR(p) and MA(q) models that are appropriate for univariate time series modeling. In the AR(p) model, the next value is simulated as linear relationship of a constant term, random error, and  $p$  previous observations (Adhikari and Agrawal, 2013). There are several statistical compositions regarding ARMA models with time-varying coefficients and with random coefficients. A well-defined ARMA model is often utilized for estimating a stationary stochastic process.

It is assumed that the coefficients  $\phi_1 \dots \phi_p$ ,  $\theta_1 \dots \theta_q$ , and  $\sigma^2$  do not depend on time. The order of  $p$  and  $q$  is predetermined, and parameters  $\phi_1 \dots \phi_p$ ,  $\theta_1 \dots \theta_q$ , and  $\sigma^2$  may be predicted from recognition using several estimation techniques. Order is determined through observation



(Choi, 2012).

### 2.4.5 Autocorrelation and Partial Autocorrelation Functions

Autocorrelation and partial autocorrelation are both time series statistical methods used to identify patterns and correlations between data in a series. Both are required for understanding the patterns of time series data and developing reliable forecasting models. In short, ACF and PACF are examined to access suitable models for time series data. Thus, these statistical measures depict the relationships between data in time series. Therefore, for modeling, ACF, and PACF plots were developed as well as predicted. The arrangement of the AR and MA words is determined by these plots (Adhikari and Agrawal, 2013). ACF and PACF associated with MA and AR are given in Table 2.1 (Anderson, 1977).

Table 2.1: Identification of AR and MA order through ACF and PACF

Process	ACF	PACF
AR(p)	Damps out	Cuts off after lag p
MA(q)	Cuts off after lag q	Damps out
ARMA(p,q)	Damps out	Damps out

### 2.4.6 Augmented Dickey-Fuller Test

The Augmented Dickey-Fuller (ADF) test is performed to determine if data collection is stationary or not. The following equation is a generalized version of this method:

$$\Delta y_t = \alpha_o + \alpha_1 y_{t-1} + \sum_{i=1}^n \alpha_i \delta y_t + \epsilon_t \quad (2.8)$$

The above equation consists of a data series  $y$  in time  $t$ , where  $n$  represents an ideal number of lags,  $\alpha_o$  represents fixed value, also  $\epsilon$  represents a white noise error (Ghauri et al., 2020).

## 2.5 Exponential smoothing State Space Model

A type of time series forecasting approach that makes forecasts using exponential smoothing techniques. A time series is decomposed into three components by the ETS model, which are: error, trend, and seasonality. ETS(A, M, N) denotes the error type (Additive or Multiplicative), M denotes the trend type (None, Additive, or Multiplicative), and N denotes the seasonality type (None, Additive, or Multiplicative).

Jain and Mallick (2017) used the basic Exponential smoothing to discrete data. Also, the approach must be extended to account for trend and seasonality. Taylor suggested first-ordered autocorrelation in residuals while using the ETS technique to predict time series. Each ETS model is a stochastic model. It is regarded as a paradigm of judicial extrapolation. The explanation in the wake of this model was the forecast of the weighted average of prior data, and weight decayed exponentially with time. Recent observations were given exponentially greater weight in the exponential smoothing model than historical observations. There were fifteen different exponential approaches that were distinguished by the seasonality and trend behavior of the data. The trend component might be additive or additive damped, multiplicative or multiplicative damped, or not present. The seasonality component could be multiplicative, additive, or not present. A triplet (ETS) is utilized to distinguish between these models.

Error, Trend, and Seasonality are abbreviated as ETS. All of these models use the general vector  $u_t = (l_t, b_t, s_t, s_{t-1}, \dots, s_{t-m+1})'$ , and the state space equations take following notation:

$$v_t = w(u_{t-1}) + r(u_{t-1})\epsilon_t \quad (2.9)$$

$$u_t = f(u_{t-1}) + g(u_{t-1})\epsilon_t \quad (2.10)$$

here  $\mu_t = w(u_{t-1})$ ,  $\epsilon_t$  is a Gaussian white noise process with mean zero and variance  $\sigma^2$ . The model with additive error has  $r(u_{t-1}) = 1$ , hence  $v_t = \mu_t + \epsilon_t$ .

The method with multiplicative biases has  $r(u_{t-1}) = 1$ , so that  $v_t = \mu_t + \epsilon_t$ . Therefore, for the multiplicative model  $\epsilon_t = (v_t - \mu_t) / \mu_t$  is a relative error and any value of  $r(u_{t-1})$  will lead to the same point prediction for  $v_t$ . To increase the efficiency, a machine learning forecast-

ing model RBFNN is used, which determines nonlinear function ([Panigrahi and Behera, 2017](#)).

## 2.6 Radial Basis Function Neural Network Model

RBFNN is a type of artificial neural network that has been used widely for various purposes, including function estimation, pattern recognition, and regression. The RBFNN was first proposed in the early 1990s, and its evolution can be traced back to a few major researchers and papers. The paper "Neural Networks for Time-Series Forecasting" by Paul J. Werbos, published in 1990, was one of the earliest and most prominent publications on RBFNNs. RBFNN defined the stimulation of every node in the concealed layer. A forecasting model was developed by combining the radial basis function (RBF) neural network with the adaptive neural fuzzy inference system (ANFIS) to examine how real-time electricity prices affect short-term load. The model employed the nonlinear approaching capacity of the RBF network to forecast, demand on the day of the prediction without using electricity price into account, and then it used the ANFIS system to adjust the load forecasting results that were achieved by the RBF network depending upon current changes in real-time price. The RBF network's issues would be resolved by this system integration, which would also improve forecasting accuracy. The results of a real forecasting case in this study showed how reliable the suggested model was ([Yun et al., 2008](#)).

The RBF neural network is a forward network model that performs well, provides accurate global approximations, and has no issues with local minima. An input layer, a hidden layer, and an output layer make up its three layers. It is a multi-input, single-output system as a result. When processing input, the hidden layer uses nonlinear transforms to extract features, whereas the output layer offers a linear combination of output weights. Fig. 2.1 illustrates this structure:

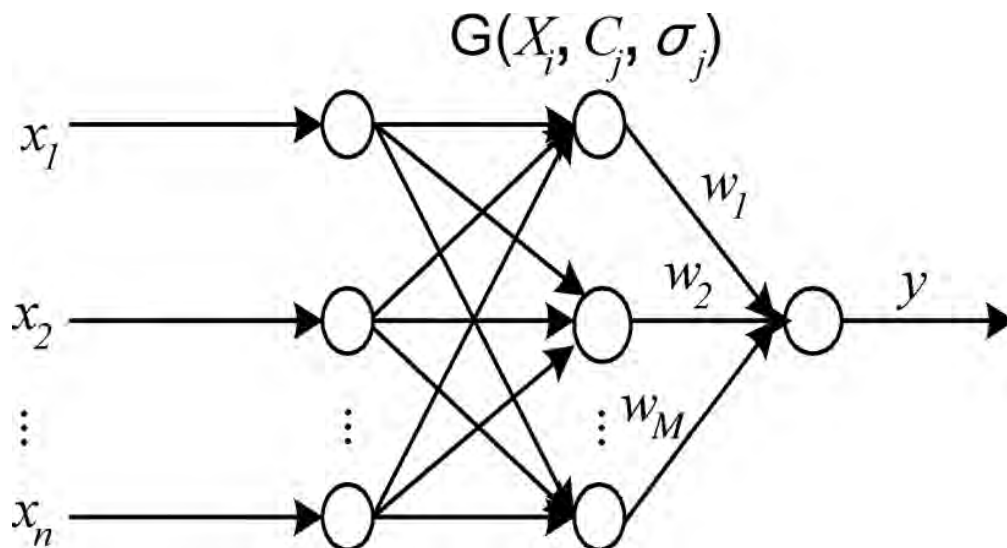


Figure 2.1: Design of Radial Basis Function Neural Network

This model estimates short-term load without taking into account the price of electricity; prediction accuracy is affected, and the outcomes are unsuitable for how the load is affected by the price of electricity. As a result, ANFIS is utilized to change the outcomes. (Yun et al., 2008).

## 2.7 Result and Discussion

Time series data is mostly dependent on time because mean and variance do not remain constant across time. The stationarity of data should be determined before starting to fit different models. There are various tests that can be utilized to evaluate if a series is stationary or not. ADF, KPSS, and PP tests are well-known stationary tests. There are also stationarity unit root tests, ADF unit root, KPSS unit root, and PP unit root tests on direct tax time series data of total tax revenues, here, the ADF test is used. Firstly, the ADF test confirms the nonstationarity of direct tax time series data. The data is non-stationary according to the null hypothesis but stationary according to the alternative hypothesis.

Table 2.2: ADF test for checking stationarity of direct tax data of time series.

Series	test-statistics	p-value
Direct Tax Revenues	-0.92759	0.9483

As a result, because the p-value is greater than 0.05, the null hypothesis cannot be rejected and concluded that the time series data is nonstationary.

Figure 2.2: Plot showing trend and seasonality in original series of direct tax revenues

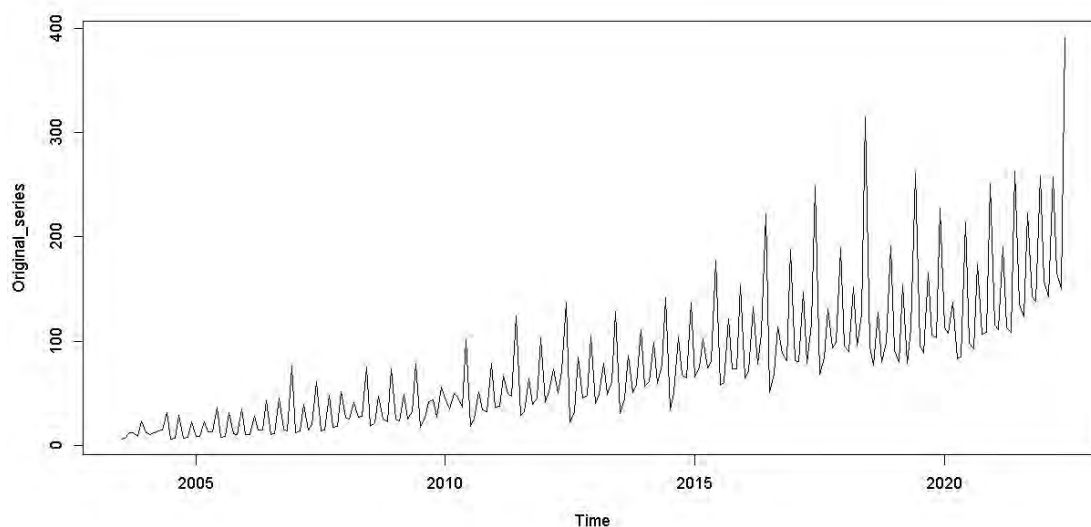


Figure 2.2 shows the data's trend as well as seasonality. Differencing is the most commonly used method for converting non-stationary data to stationary data by subtracting observations from one another. The mean of the time series is stabilized through differentiation by eliminating variations in the ranks of time series and therefore lowering trend and seasonality. Here, differencing is used to transform non-stationary data into stationary data.

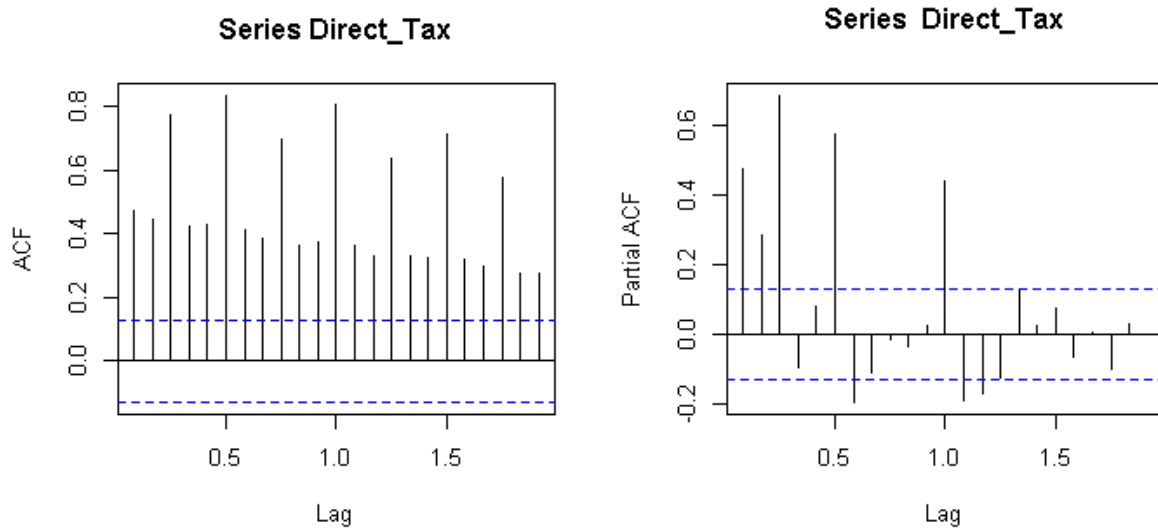
## 2.8 Identifying Model parameters

Firstly, by watching PACF and ACF, the parameters of AR( $p$ ) and MA( $q$ ) can be determined. ACF and PACF are commonly used to determine the order of an MA model AR model. ACF refers to simple correlation with its value, also known as serial correlation, whereas PACF refers to the amount of mutual correlation after removing the effect of other variables.

The AR model's parameter is  $p$ . The order of AR terms can be established by the PACF cuts off after lag  $p$ , and similarly for MA terms by the ACF cuts off after lag  $q$ , where  $d$  is for difference data, and because we only take one difference, after the first differencing it becomes stationary data from the number of differencing values of  $d$  has been determined. To apply

the above theory, ACF and PACF of the series are provided below in the figure

Figure 2.3: Correlogram of series for Direct Tax time series data

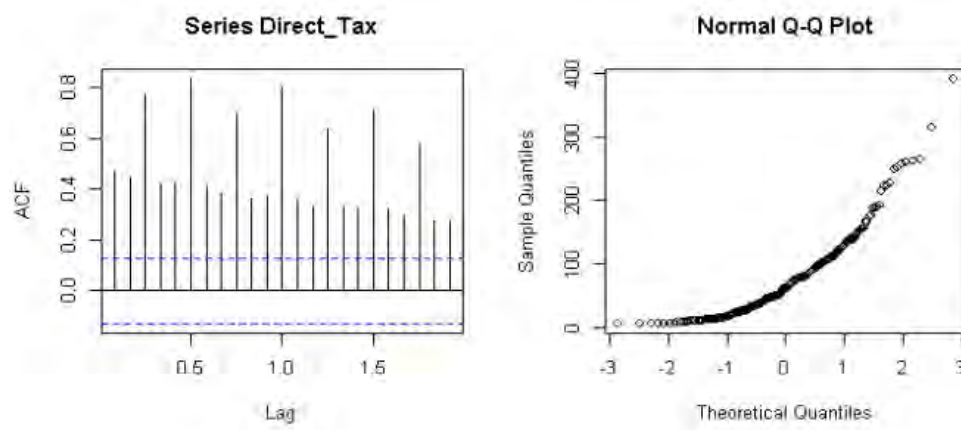


### Diagnostic checks

The ARIMA model for direct tax revenues is also developed. ARIMA is a simpler model for forecasting future values in a series. The ARIMA model is determined using Maximum Likelihood Estimation (MLE). The model's three key parameters are  $p$ ,  $d$ , and  $q$ . Seasonal ARIMA is an improvement over ARIMA.  $S$  is the seasonal periodic behavior period, which is 12 because the series is monthly data. It is written as ARIMA( $p$ ,  $d$ ,  $q$ ) ( $P$ ,  $D$ ,  $Q$ ) $S$ . The best model generated in this series is the ARIMA(1,0,2)(0,1,0) [12]. In it, the seasonal parameter ( $P$ ,  $D$ ,  $Q$ ) is (1,0,2), where 12 is the seasonal component with non-seasonal ARIMA (0,1,0). Investigating the result of the best fit ARIMA(1,0,2) can be written as

$$X_t = Y_t - 0.4647Y_{t-1} + \epsilon_t \quad (2.11)$$

A diagnostic check is subjected to the direct tax data shown in Figure 2.7.

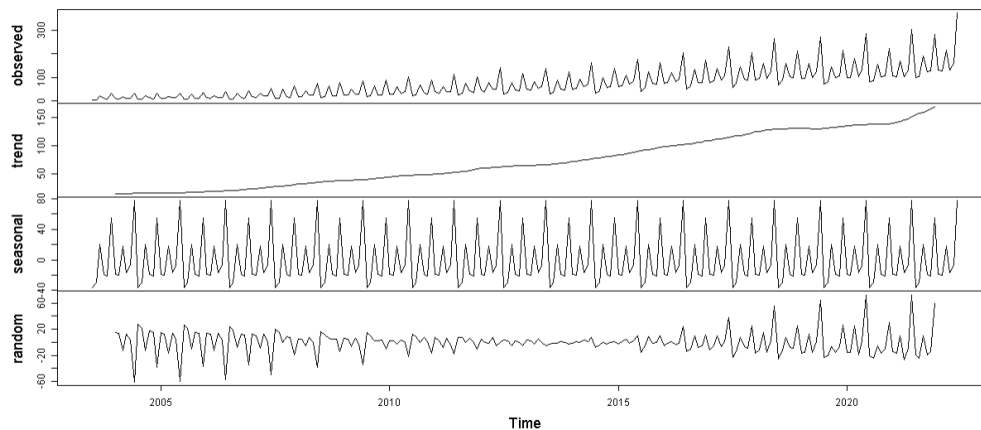


Above Fig shows ARIMA(1,0,2) model fulfills all assumptions, residuals are normally and independently distributed. Hence, as a result, it is ready for forecasting.

The other model used in the series is the ETS model. The ETS is responsible for automatically selecting an appropriate model based on the maximum likelihood method in order to forecast numerous steps ahead as needed. The simplest model from 30 provided state space models is chosen based on AIC, BIC, and AICc. The ETS generates the best model in this series is ETS(M, A, M). The series' best model is generated by ETS(M, Ad, M). This three-character string in ETS represents error, trend, and seasonality. An error can be additive or multiplicative, a trend can be additive, additive damped, multiplicative, multiplicative damped, or none, and a seasonality can be additive, multiplicative, or none.

ETS(M, Ad, M) is the best model among the others. It denotes multiplicative error and seasonality, which suggest that the parameter is amplified over time, whereas the trend is additively damped, which means that the size of the trend is diminishing over time and approaching a straight line. There are four parameters i.e.,  $\alpha$ ,  $\beta$ ,  $\phi$  and  $\gamma$  where  $\alpha$  is smoothing parameter for level,  $\beta$  is smoothing parameter for trend,  $\phi$  is damping coefficient and  $\gamma$  is smoothing factor for seasonality. Small value of  $\beta = 0.0012$  and  $\gamma = 0.2817$ . The value of  $\beta$  suggests that recent observations are given very little weight in the trend component, indicating a rather slow-changing trend. Similarly, the value of  $\gamma$  indicates that the seasonal component is relatively responsive to current observations, implying that the model takes recent seasonality patterns into account while making forecasts. Hence, both values show trend and seasonal components do not vary rapidly with time. Specifically,  $\ell$  represents the level equation,  $b$  represents the trend equation, and  $s$  represents the seasonal equation.

Figure 2.4: Plot for best ETS model, showing observe value and slope along with trend and seasonality component:



The Radial Basis Function Neural Network (RBFNN) is a type of artificial neural network that employs radial basis functions as activation functions. RBFNN builds short-term forecasting through RBFNN-type neural network algorithms. RBFNN determines the relation among lags of the series of direct taxes. RBFNN is different from the well-organized method used for forecasting.

Each lag refers to a previous observation in the series. Its ability to capture the connection between these lagged values allows it to learn and predict future time series values. RBFNNs can be seen as a collection of algorithms and techniques working together to solve different types of problems, particularly those requiring pattern recognition, function approximation, and time series forecasting.

Table 2.3: RBFNN showing short-term forecasting of direct tax

Month	Point forecast	Lo 95	Hi 95
Jul 2022	138.0375	84.98292	191.0921
Aug 2022	131.3897	80.64238	182.1370
Sep 2022	131.3897	146.64129	332.6994
Oct 2022	151.4259	92.35369	210.4980
Nov 2022	152.1060	92.46664	211.7453

Table 2.3 shows forecasting points after the Jun 2022 observation, giving the next predicted six observations. The predicted amount of direct tax for Jun 2022 is 138.0375. For the



95% confidence interval, direct tax is 84.98292 to 191.0921. Similarly, Table 2.3 displays the same information for the remaining four observations. The inaccuracy of the estimated and the data set's observed values can be utilized to compute the model's performance. The performance of the measure is evaluated by computing the Mean square Error (MSE) for all models, as shown in table 2.4.

Table 2.4: Comparison of mean square error of ARIMA, ETS, and RBFNN model

Models	model	Mean Squared Error
ARIMA	one stage	9612.198
<b>ETS</b>	<b>one stage</b>	<b>8665.726</b>
RBFNN	one stage	9648.13

Hence, the effectiveness of ARIMA, ETS, and RBFNN for direct tax data on total tax revenues with a minimum value of MSE is shown in Table 2.4. The ETS model performed more accurately and more consistently. Among all other models, ETS(M, A, M) produces the best outcomes. Many hybrid models are proposed in the next part to overcome the drawbacks of process-based models discussed above and to produce more efficient results.

## 2.9 Conclusion

The goal of univariate time series modeling is to estimate data where each observation is recorded once at regular intervals. The ARIMA model, "AutoRegressive Integrated Moving Average," is frequently used to interpret this data. Complex linear relationships in time series data are extremely well captured by ARIMA model.

ARIMA, ETS, and RBFNN are the three competitors for a single stage. When processing time series data, ARIMA and ETS have a lot of flexibility. ETS models are particularly useful when the error, trend, and seasonality components of the data are obvious. Three divisions of the time series make it simpler to comprehend and forecast future values. ETS models are resilient even when the data exhibits patterns but fall short of the tight linearity requirements of ARIMA. They may deal with time series that vary in complexity and seasonality.

RBFNN, also known as the Radial Basis Function Neural Network, is the third competitor.

For tasks like approximating functions and identifying patterns, machine learning utilizes radial basis functions in its hidden layers. ETS minimizing Mean Squared Error (MSE) 0.185.1431 in just one stage. Time series data with obvious linear dependencies and stationary patterns can be handled well by ARIMA models. They are frequently utilized for predicting, particularly when the data are dominated by these linear correlations.

At one stage here, ARIMA provided a result of 9612.198. That may appear to be a large figure, but what it actually represents is the Mean Squared Error (MSE) of our predictions. It's a gauge of how well the ARIMA model projected the time series data, to put it simply. A larger MSE, like the one obtained with ARIMA, indicates that there was a little more inaccuracy in the study forecasts than there was in the actual data. Therefore, even though ARIMA is a strong tool for modeling time series data, it fell short in this case in terms of accuracy for this one-stage forecast. Let's now discuss ETS, which at one point provided an impressively low MSE of 185.1431. This low value is encouraging because it shows how well the ETS model forecasts the time series data. In essence, it means that the ETS forecasts were quite accurate in predicting the data points. Dealing with time series data that exhibits distinct patterns, such as trends and seasonality, is where ETS excels. It separates the data into various parts and uses that knowledge to analyze outcomes a high level of precision. In this instance, ETS's forecasting abilities were outstanding. We have RBFNN, which at one point generated an MSE of 9648.18. This outcome is more in line with observed ARIMA. In essence, compared to the real data, RBFNN's predictions had a little bit higher error. In conclusion, these findings help in understanding how each modeling strategy performed when it comes to the presentation of future values of this time series data. Hence, ETS showed high accuracy in this one-stage prediction due to its reduced MSE. Although still useful tools, ARIMA and RBFNN displayed a little bit higher prediction error in this particular situation. It's vital to select the modeling strategy that best matches the properties of data and forecasting objectives.

# Chapter 3

## Decomposition Techniques and Hybrid Models

### 3.1 Introduction

Wavelet Analysis (WA) and Empirical Mode Decomposition (EMD) are two distinct signal processing techniques employed for different purposes, including denoising and time series data analysis. Wavelet Analysis involves decomposing a signal into its component wavelets at various sizes. Wavelets can capture both the high-frequency and low-frequency elements of signal since they are essentially little waves that change in frequency and amplitude. WA is frequently used for denoising due to its ability to distinguish between signal and noise components. It is feasible to filter out noise while keeping the signal's important characteristics by figuring out the scales or frequencies at which the noise predominates. WA can capture both localized and global aspects because the data is represented on several scales. The signal-to-noise ratio can be improved by properly separating signals from noise using WA. There are numerous applications, including image processing, data compression, and pattern recognition. It can be used for feature extraction. Empirical Mode Decomposition (EMD) is a data-driven signal processing approach employed to decompose complex time series data into a collection of intrinsic mode functions (IMFs). According to frequency and amplitude, each IMF reflects a different oscillatory component of the data. EMD is especially helpful when working with non-stationary and non-linear time series data, which cannot be suitable for more conventional techniques like Fourier analysis. Extraction of IMFs from the original

signal iteratively is a key component of EMD. Local peaks and minima are used to determine IMFs, which are then isolated by computing the envelope of the signal and subtracting it from the signal. EMD can be used to decompose a noisy signal and denoise it into its constituent IMFs and then reconstruct the signal using only those IMFs that accurately represent the important signal components; EMD can be used to denoise signals. In summary, both WA and EMD are powerful techniques for denoising and examining time series data. While EMD is particularly useful for managing non-linear and non-stationary data, making it ideal for diverse applications where standard approaches may fall short, WA is used for its multiscale analysis and noise reduction capabilities.

Time series forecasting is a prominent modeling method that involves the analysis of past values about related variables. In their study, [Wang et al. \(2016\)](#) examined the comparative effectiveness of hybrid models in predicting short-term traffic conditions as compared to conventional models. The researchers aimed to investigate the dynamic behavior of hybrid models under varying circumstances and elucidate the reasons behind their superior forecasting capabilities in comparison to single-stage models. The predictive outcomes of the EMD-ARIMA model contrasted with those of the ARIMA, Holt Winter, Artificial Neural Network (ANN), and a naïve model. Based on the results, it appeared that the hybrid EMD-ARIMA model performed better than other single-stage models.

Wavelet analysis (WA) applications in hydrology were examined by [Sang \(2013\)](#) also, a summary was provided. The author identified six key applications of WA, including wavelet-aided cross-correlation analysis of hydrological series, wavelet-aided hydrological series simulation, and wavelet-aided hydrological series forecasting. These applications include wavelet-aided multi-temporal scale analysis, wavelet-aided deterministic component identification, hydrological series denoising with wavelets, and wavelet-aided complexity quantification of hydrological series. The results showed that continuous wavelet analysis and wavelet spectrum analysis were emphasized most in wavelet analysis. and hydrological series supported by wavelets. The use of wavelet-denoising, wavelet-aided complexity description, and wavelet cross-correlation analysis, however, is not given much prominence. Future advancements in its use might result from a clearer comprehension of these three factors. This analysis clarified the hybrid methodology's better performance versus a single method. In subsequent iterations, this

approach had the potential to be enhanced to an admirable degree.

[Kim and Oh \(2009\)](#) presented a novel R package that offers a sound application of Empirical Mode Decomposition (EMD) and Hilbert spectrum characterizing signals affected by noise. The EMD technique employs a multi-resolution instrument that utilizes intrinsic mode functions (IMFs) also spectral analysis for capturing time-varying amplitudes and stages in a scale-dependent manner. Enhancements might be achieved by the actual use of the Empirical Mode Decomposition (EMD) approach across several statistical applications. [Rabbani et al. \(2011\)](#) explained the advantages of using Hilbert and wavelet transforms, as well as changing the threshold approach, to provide an optimal combinational strategy for the identification of R peaks and achieve favorable outcomes. The proposed approach used a combinational strategy to the degree of sensitivity of R as a peak determination method in noise presence. The findings indicated that all three strategies had a significant impact on the detection of the R wave. The efficacy of the approach might be enhanced by using transforms that provide a superior energy compactness quality, such as the complex wavelet transform.

[Duraivel et al. \(2020\)](#) went into the intriguing field of finger movement-related brain signal decoding. The main purpose of this study was to evaluate the efficacy of four different strategies for collecting neural data: Fast Fourier Transform bandpass filtering (FFT), Principal Spectral Component Analysis (PSCA), Wavelet Analysis (WA), and Empirical Mode Decomposition (EMD). Duraivel and his team's goal in this scientific investigation was to determine how well each of these tools—FFT, PSCA, WA, and EMD—performed on the basis of giving insightful knowledge about the neurological data. Comparing different detectives working on the same case is similar since they all have different investigation philosophies. They found that not all tools were created equally. Some techniques were much more successful than others at revealing important details about finger movements. Researchers and engineers working on neural decoding pipelines will find this knowledge to be a gold mine. It's similar to understanding the best tools to use to decipher a particularly challenging code. Better neuroprosthetics, a deeper understanding of brain diseases, and improved brain-computer interfaces are all on the horizon.

### 3.1.1 Continuous Wavelet Transform

In CWT, the mother wavelet  $\Phi$  has been transformed into a family of "wavelet daughters"  $\Phi(y, z)$  by translating and scaling. Assume that a time series  $Y_i$  has a continuous scale and a discrete pattern where  $t = 0, \dots, t - 1$ , and that the wavelet function  $\Phi$  lean on  $t$  is typically defined as:

$$CWT(y, z) = \frac{1}{\sqrt{y}} \psi\left(\frac{t - z}{y}\right) \quad (3.1)$$

$$W_{\psi} f_{y,z} = \int_{-\infty}^{+\infty} Y(t) \psi_{y,z}^*(t) dt = \frac{1}{\sqrt{y}} \int_{-\infty}^{+\infty} Y(t) \psi^*\left(\frac{t - z}{y}\right) dt \quad (3.2)$$

In a CWT plot, frequency is shown as a function of time, while time is shown as a function of frequency. The height of a scale and how it changes over time may be determined via a scalogram if the scale  $y$  and translation  $z$  functions are uniformly varied with time ([Chaovalit et al., 2011](#)). A wavelet's position in the frequency domain is denoted by  $y$ , while its position in time is indicated by  $z$ . Time and frequency information may be gleaned by identifying the real series as a function of  $z$  and  $y$  ([Soares and Aguiar-Conraria, 2006](#)).

### 3.1.2 Discrete Wavelet Transform

The DWT function was proposed by Mallat in 1989 with its own unique band-pass and low-pass properties for each mother wavelet to break down a signal using a program of modulation mirror decomposition filters ([González-Audicana et al., 2005](#)). Since wavelet analysis didn't needlessly repeat the information already there in the coefficients of the wavelets, the full original signal may be reborn. In order to reconstruct an ECG signal with high quality, DWT filters out background noise and provides a high compression ratio ([Olkkonen, 2011](#)). When it comes to practical methods and signal reconstruction, the DWT is all that's needed. It provides enough information and makes a reasonable suggestion to speed up the computing process. Noise filtering is a key use of DWT. Experts in the area often define noise as significant outliers in the data that significantly distort the original signal (Han and Kamber 2006;

Orfanidis 1996). Noise filtering is based on the principle of isolating relevant information from background noise. Therefore, a suitable noise filtering technique should be able to eliminate the noise and isolate the signal. DWT is often used to investigate time series data that is neither stationary nor linear. Time series data is transformed using a technique called wavelets. The modification results in a smaller series and less background noise. DWT is employed for decomposing time series into a variety of scales and frequencies. Challenges to DWT include concerns like wavelet selection, analysis depth, boundary problems, and data dependence (since DWT is a data-dependent technique). DWT is not very helpful and significant in comparison to other methods if the data is stationary or when there is no trend in the series being examined. EMD is a non-stationary data decomposition technique (Chaovalit et al., 2011).

## 3.2 Techniques of Decomposition

Wavelet Analysis (WA) and Empirical Mode Decomposition (EMD) are both time series decomposition techniques that include dividing time series in its basic elements i.e., trend, seasonality and noise. Here are their advantages and disadvantages: WA is especially good at capturing information at various scales or levels. As a result, it can be used to find patterns and components at various degrees of depth within a time series. Wavelets excel at time-frequency localization, allowing for the exact identification of transitory characteristics and abrupt changes in a time series. Wavelet analysis can be used for a variety of data sources and is appropriate for both stationary and non-stationary time series. WA can provide insights into the frequency domain features of a time series, assisting in the effective identification of periodic components.

The wavelet function and its parameters can have a major impact on the results. On the other hand, EMD is a data-driven and adaptable technique that has the ability to adapt the inherent properties of the data, making it suited for a wide range of time series, including non-stationary and nonlinear data. Unlike some different decomposition techniques, EMD does not require the selection of basis functions in advance. It derives basis functions directly from data, which can be useful in various situations. EMD excels in capturing nonlinear and non-stationary time series components, making it valuable for analyzing difficult data.

EMD provides a series of intrinsic mode functions (IMFs) that are simple to comprehend and can reveal hidden patterns in data. EMD may be affected by mode mixing, which occurs when the decomposed IMFs contain mixed components, making it difficult to comprehend their physical significance. EMD can be computationally demanding, especially for large time series, which may limit its utility in some applications.

Hence, both WA and EMD have distinct advantages and disadvantages in time series decomposition. The two methods should be chosen based on the specific qualities of the data and the goals of the research. WA captures patterns at various scales and provides improved time-frequency localization, whereas EMD is adaptive and especially well-suited for nonlinear and non-stationary data.

### 3.2.1 Empirical Mode Decomposition

Rapid development has occurred in several scientific and engineering disciplines by utilizing concept of EMD. [Huang et al. \(1998\)](#) were the first to develop EMD. Its primary function is to separate a signal into its individual IMF components ([Kim and Oh, 2009](#)). It is applicable to information dominated by psychological considerations. Data development in terms of intrinsic mode functions (IMFs) lies at the heart of the EMD, making it a data-driven technique. It demonstrates a non-linear process when applied to a non-stationary series. Unlike signal management techniques, EMD is unique ([Singh and Borah, 2014](#)). It provides precise, data-driven signal breaking for both fast and slow oscillations. At the end of the day, the true signal may be broken down into a set of "intrinsic mode functions" governed by amplitude and frequency. EMD may be used as a pass filter, much like a rectifier. It begins by converting the fast-moving parts of a complicated signal. One of EMD's primary functions is the removal of IMFs ([Rabbani et al., 2011](#)).

### 3.2.2 Intrinsic Mode Functions

[Kim and Oh \(2009\)](#) stated that identifying a signal's oscillation at the local time scale is the first step in picking IMFs. The data-determined IMFs are then used to provide a starting point for an extension that may or may not be linear. There are many conditions that have



been met by IMFs demonstrating the data's inherent time scale.

1. The number of zero points and extrema are substantially identical or different by no more than one.
2. The mean value of the case is zero at any point according to local extrema and minima. The frequencies IMF1 contains are highest, while frequencies in the residual are the lowest (Rabbani et al., 2011). IMFs are a crucial tool for removing noise since they function as a filter bank.

### 3.2.3 Steps of Decomposition into Intrinsic Mode Functions

Decomposing a signal into IMFs involves the following steps:

1. The real signal of  $x(t)$ 's positive (maxima) and negative (minima) apexes must first be determined.
2. Create the signal's minimum  $w(t)$  and maximum  $z(t)$  envelopes using the cubic spline method.
3. Determine the mean by averaging the upper and lower envelopes.

$$n(t) = (w(t) + z(t))/2 \quad (3.3)$$

4. To create a first detailed signal while subtracting the average of the original signal's envelopes.

$$d_1(t) = x(t) - n(t) \quad (3.4)$$

5. Untill the detail signal  $d_k(t)$  meets the above two circumstances of IMF  $k_1(t) = d_k(t)$ , iterate the above 1 – 4 steps,till the IMF of  $x(t)$  is obtained.
6. To get all IMFs, iterate 1 – 5 steps on each subsequent residual  $h_n(t) = x(t) - k_n(t)$  to obtain all IMFs.

The original series is shown below:

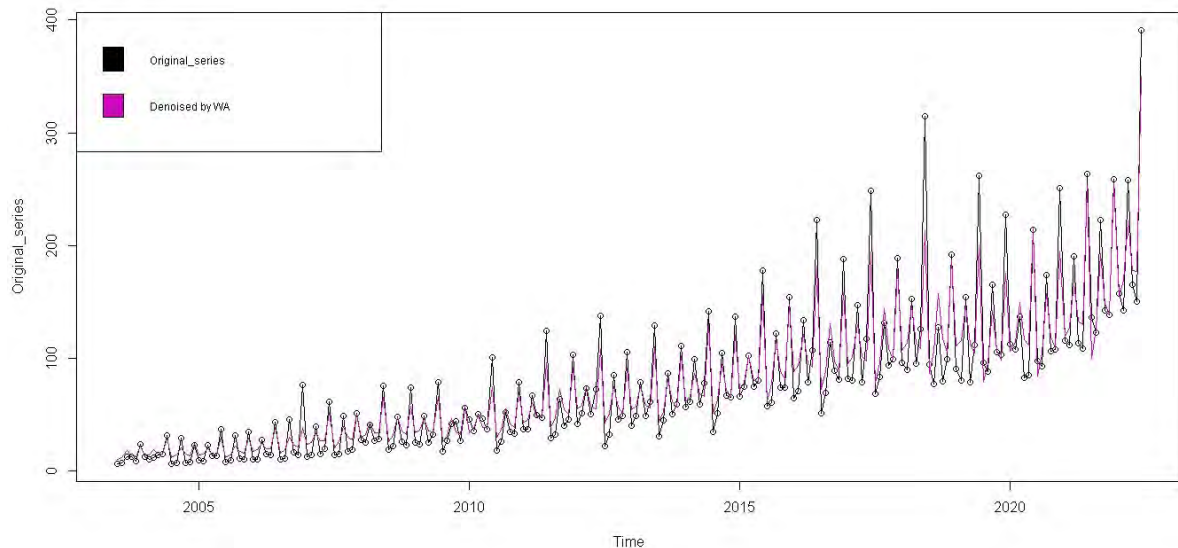
$$x(t) = \sum_i^n k(i) + h(t) \quad (3.5)$$

In EMD, mode mixing is a critical issue. To address this issue, a suitable CEEMDAN approach was used (Kim et al., 2021).

### 3.3 Results and Discussion

The stationarity of the series has been checked by stationary test, i.e., ADF test in the previous section, which shows the series is non-stationary. To remove noise from the series, DWT is used to denoise the data of direct tax. In this study, a graph is created using Wavelet analysis (WA). For this, use a two-stage hybrid model in which the series is denoised first. Soft and hard thresholds are used to make the series smoother, and wavelet-based models WA-ARIMA, WA-ETS, as well as WA-RBFNN are developed. Prediction models are combined with the machine learning-based model RBFNN for comparison with traditional statistical models. The denoised series of direct tax is shown in the Fig 3.1 :

Figure 3.1: The Denoised series of the direct tax obtained by WA(in purple color)



In the previous section, a one-stage model was used, now this study proceeded to generate a graph using Wavelet analysis. For this use a two-stage hybrid model in which the series is denoised first. Soft and hard thresholds are used to make the series smoother. Following the denoising step, the study proceeds to develop wavelet analysis-based models, specifically WA-ARIMA, WA-ETS, and WA-RBFNN. These models incorporate wavelet analysis techniques

to capture and analyze the data's frequency components and patterns. The prediction models are combined with the machine learning-based model RBFNN for comparison with traditional statistical models. WA-ARIMA represents the performance of the two-stage hybrid model in terms of forecasting or modeling the time series. This approach, which combines wavelet analysis with the ARIMA model, is likely to provide valuable insights into the dataset's behavior and may potentially outperform traditional statistical models. WA-ARIMA gives the following result shown in Table 3.1.

Table 3.1: WA-ARIMA model for the direct taxes

WA-ARIMA(3,1,3)			
Coefficients:			
	MA1	MA2	MA3
	0.3889	-0.0788	-0.7859
s.e	0.0317	0.0233	0.0181

so it shows WA-ARIMA(3,1,3) model as a best model giving MSE equal to 127.8553. The equation is shown below:

$$Y_t = Z_t - 0.3889Z_{t-1} - 0.0788Z_{t-2} - 0.7859Z_{t-3} \quad (3.6)$$

WA is applied to the non-stationary series to reduce noise, and then a threshold is used to make it more smooth. Then, ETS is used to create a hybrid model, and WA-ETS yields the outcomes shown in Table 3.2

Table 3.2: WA-ETS model showing additive error, additive damped trend and error, trend parameters

<b>WA-ETS(A,N,N)</b>
<b>Smoothing parameters:</b>
alpha = 0.2665
<b>Initial states:</b>
l = 0.0089

So WA-ETS gives the best model WA-ETS(A, N, N), it comes up with an error term. However, it lacks a pattern or seasonal component. This may be appropriate for data that fails to show long-term trends or recurrent patterns. RBFNN creates, authorizes, and selects several networks of increasing complexity until an ideal model is found. First, the wavelet transform is combined with RBFNN to model direct tax data and combine the benefits of diverse wavelets for forecasting. Basically, a process involving the use of RBFNN and wavelet transform to build and select neural network models of varying complexity for forecasting direct tax data. This approach capitalizes on the benefits of wavelet-based analysis for enhancing the accuracy and effectiveness of the forecasting process.

Table 3.3: The evaluation of the prediction error of suggested WA-ARIMA, WA-ETS, WA-RBFNN in comparison with one-stage models

Model Name	Model	MSE
ETS	one stage	8665.726
ARIMA	one stage	9612.198
RBFNN	one stage	9648.13
<b>WA-ARIMA</b>	<b>two stage</b>	<b>127.8553</b>
WA-ETS	two stage	164.4816
WA-RBFNN	two stage	2077.513

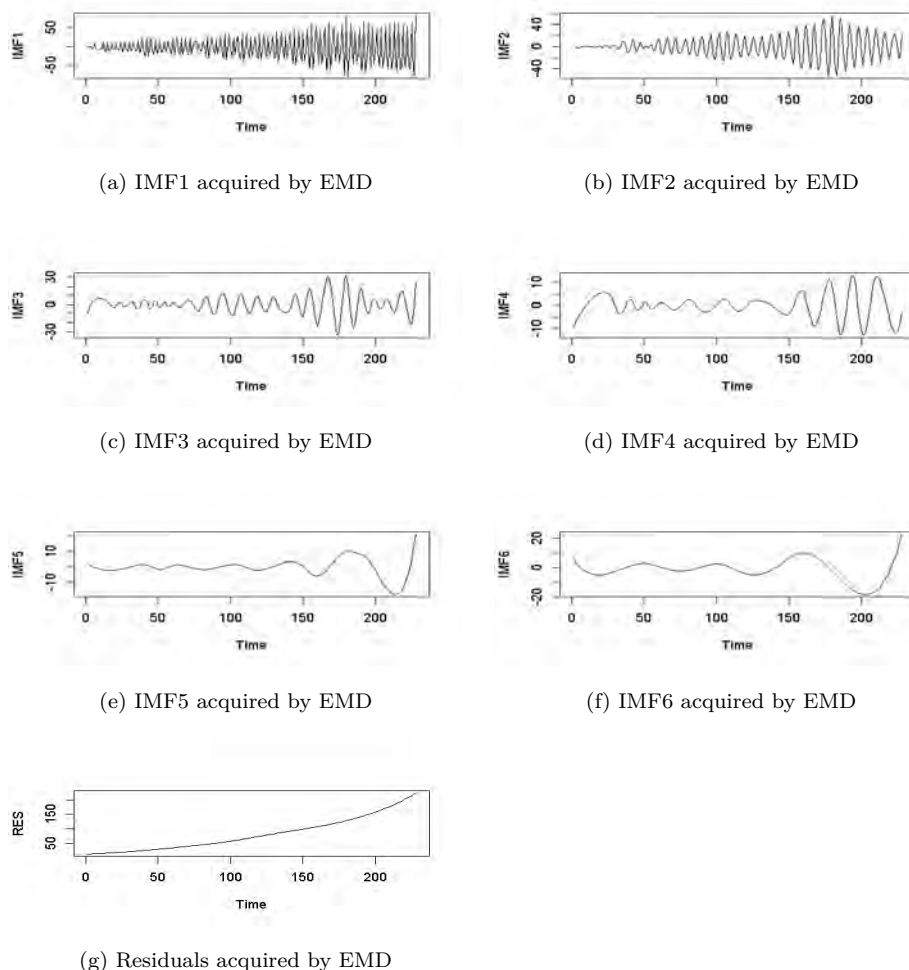
The model's performance has been evaluated by comparing the MSE of all two-stage hybrid models. Outcomes in Table 3.3 revealed the effectiveness of presented models for direct tax revenues with a minimum value of MSE in comparison to all other models. WA-ARIMA,

WA-ETS, and WA-RBFNN are the three models that have been proposed. The outcomes of WA-ETS are 164.4816, WA-ARIMA is 127.8553, and WA-RBFNN is 2077.513. From these outcomes, it is proved that WA-ARIMA performs better than WA-RBFNN and WA-ETS in terms of efficiency. The prediction performance of ARIMA and RBFNN of one stage is worse than other prediction models in predicting the direct tax of time series data.

### Decomposition using EMD

Following DWT, EMD has been used to eliminate noise from the series of direct taxes. The series decomposed into six IMFs and one residual plot of IMFs shown in Figure 3.2

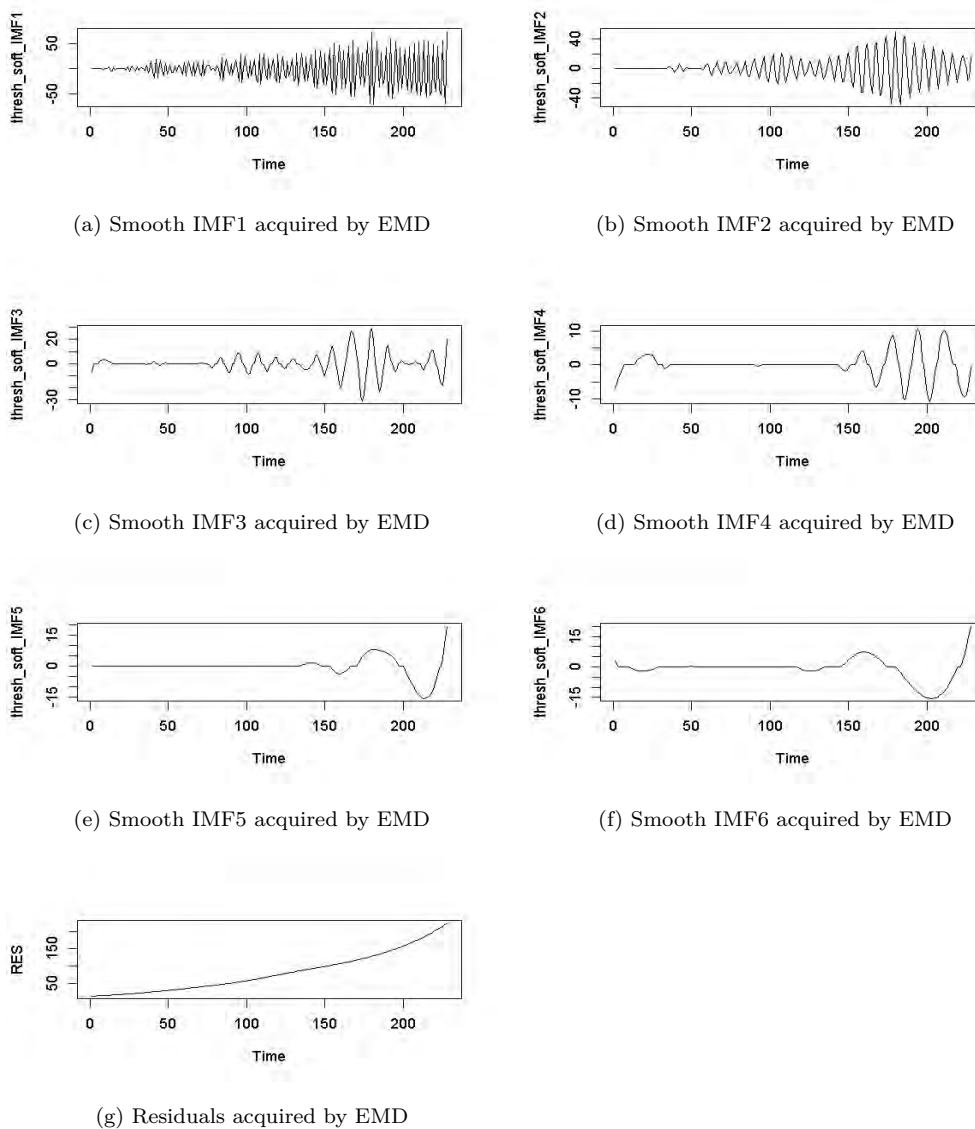
Figure 3.2: The EMD decomposition of the direct tax into six IMFs and one residual.



Soft thresholds are applied to all IMFs to decrease noise and make them smoother. Because six IMFs and one residual have a lower frequency and are practically noise-free, there is no need to apply a threshold to it because it is noise-free and has a lower frequency. The series

is recreated when a soft threshold is applied. Figure 3.3 depicts the IMFs and one residual after applying a soft threshold.

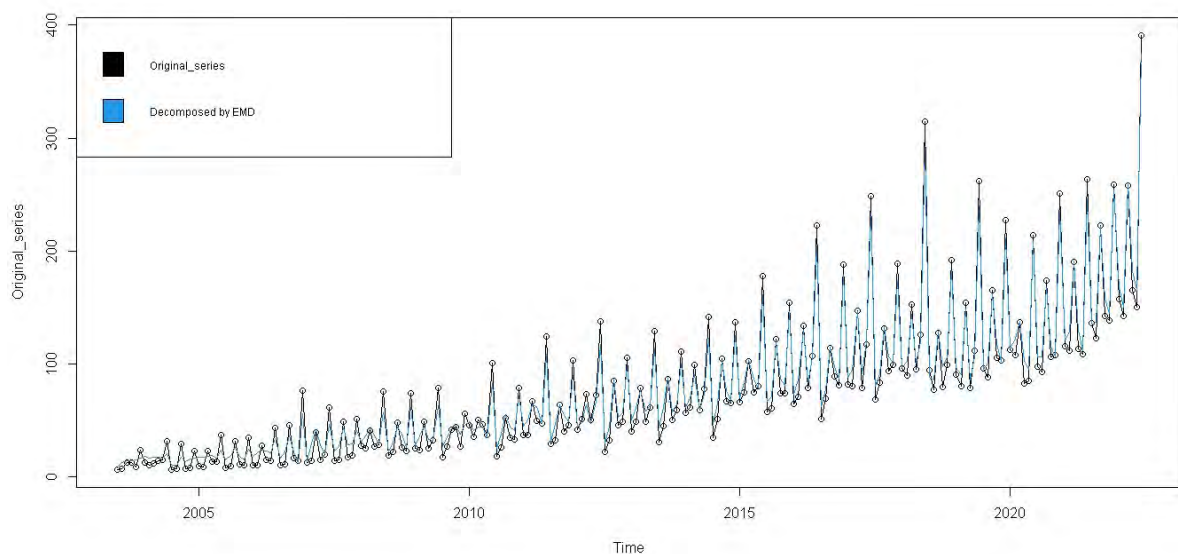
Figure 3.3: After applying the threshold, the smooth EMD decomposition of the direct tax into six IMFs and one residual.



The above series is recreated when a soft threshold is applied. Soft threshold, a valuable technique in data preprocessing, typically involves reducing the magnitude of values that fall below a certain threshold while modifying the values that exceed the threshold. In Fig 3.3, soft thresholding effectively smooths out and simplifies the data by removing or reducing high-frequency noise. Also suppressing small or noisy oscillations in the IMFs, soft thresholding simplifies the representation of the data and makes it easier to interpret and

model. Soft thresholding highlights and emphasizes significant features in the data while suppressing less important details. This can help in feature extraction for further analysis or machine learning tasks. Hence, the resulting smoothed series after soft thresholding may be more visually interpretable and can provide clearer insights into the underlying trends and patterns.

Figure 3.4: The reconstructed series, the decomposed output obtained through EMD with threshold method (in blue color)

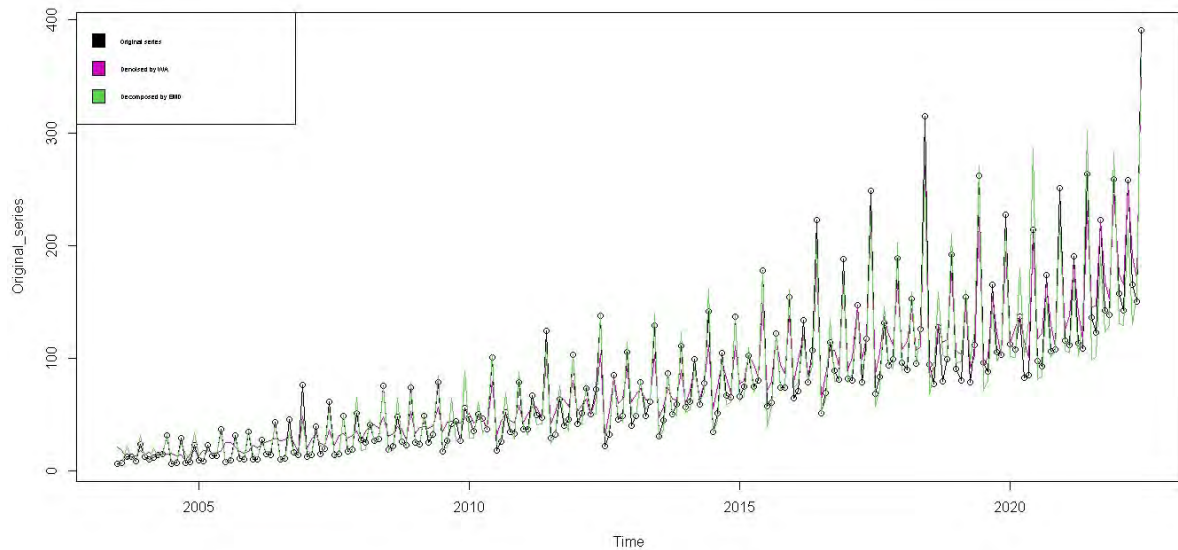


The reconstructed series is shown in the Figure 3.4. After creating the graph for the series denoised by Wavelet analysis, this study now proceeded to generate a graph using Empirical Mode Decomposition (EMD). For this use a two-stage hybrid model in which the series is decomposed first. Soft and hard thresholds are used to make the series smoother, and empirical mode decomposition-based models EMD-ARIMA, EMD-ETS, and EMD-RBFNN are developed. The prediction models are combined with the machine learning-based model RBFNN for comparison with traditional statistical models.

The original series, the Wavelet Analysis graph, and the Empirical Mode Decomposition graph are all displayed together to provide a thorough understanding of the properties of the data. While the EMD graph decomposes the series into its intrinsic modes and reveals the series' fundamental oscillatory patterns and trends, the Wavelet Analysis graph emphasizes frequency-domain features and variations at various scales. The plot for WA and EMD-based

denoised series is obtained, which is shown in Figure 3.5

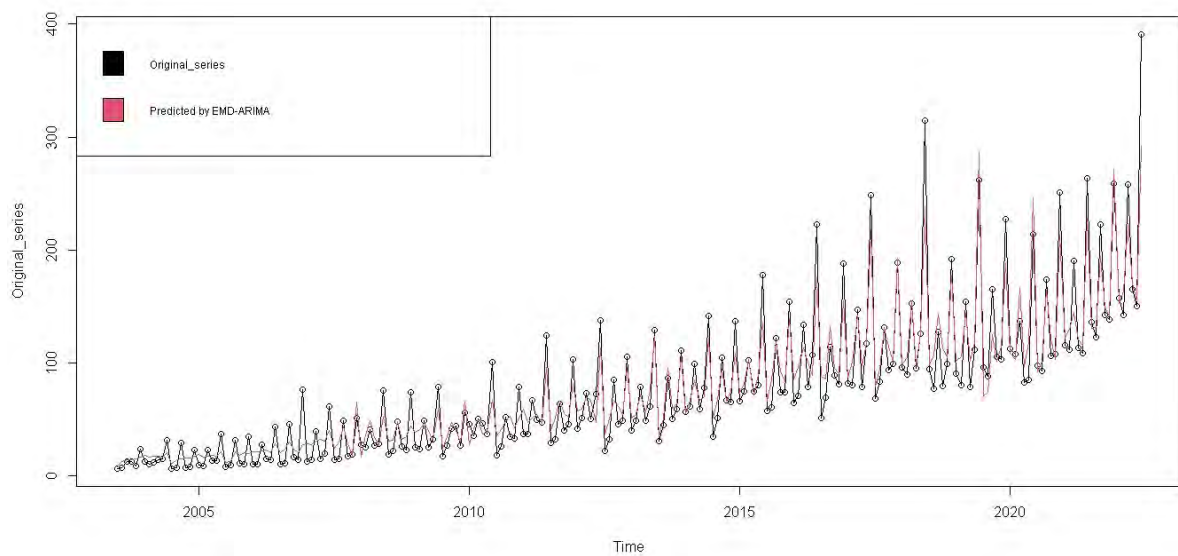
Figure 3.5: The denoised series for direct tax. The denoised is obtained by wavelet (in purple color) and EMD (in green color)



**Decompose-stage result** The ARIMA model is applied to the reconstructed data to create a hybrid model EMD-ARIMA, which yields the best model as the  $ARIMA(0,1,1)(0,1,0)$  [12] model. The hybrid model EMD-ARIMA is divided into two sections. 1) ARIMA and 2) EMD. The EMD original series is decomposed within this hybrid model. It focuses on nonstationary problems particularly. Figure 3.6. depicts the anticipated graph of EMD-ARIMA in contrast to the original data.



Figure 3.6: Prediction results of original series of direct tax in comparison with EMD-ARIMA (in red color)



Results of MSE obtained from one-stage ARIMA are worse than the EMD-ARIMA, as first decomposing and then applying ARIMA give better results than one-stage ARIMA. Therefore, compared to a straightforward one-stage ARIMA model, the two-stage EMD-ARIMA approach can be more successful when dealing with time series data that demonstrate complex, non-stationary, or multiscale patterns because it enables better modeling of the individual components before combining them to forecast the original series.

EMD-ETS is applied to the reconstructed series, which gives the best model ETS(M, A, M). It demonstrates error term is multiplicative, the trend is additive, and the seasonality is multiplicative. Predicted graph of EMD-ETS in comparison with the original series is shown in Figure 3.7.

Figure 3.7: Prediction results of original series of direct tax in comparison with EMD-ETS (in green color)

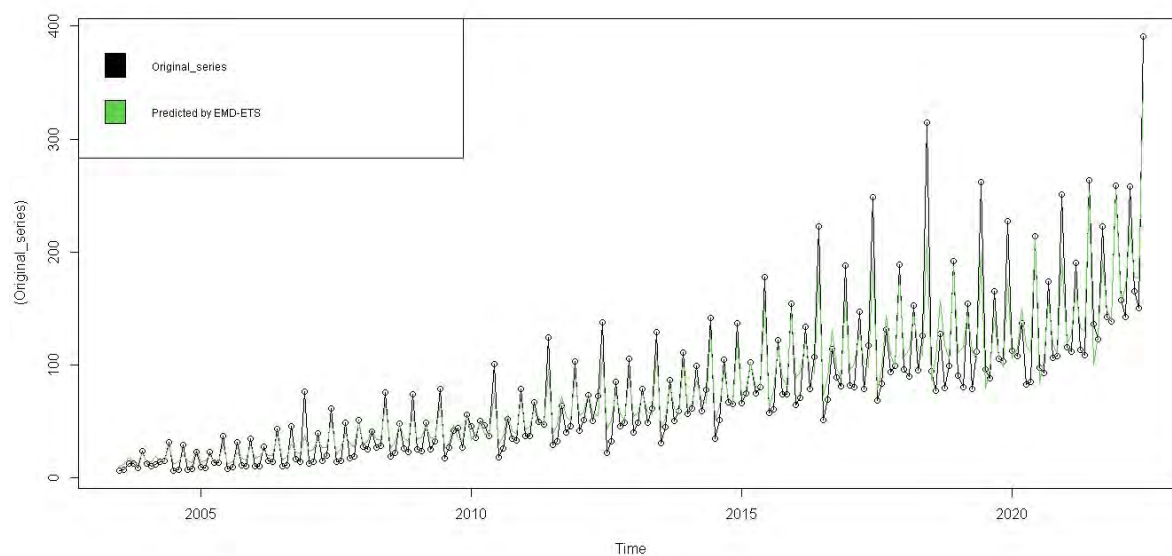


Figure 3.7 shows the graph of EMD-ETS with the original series of Direct Tax. EMD-ARIMA performs better than EMD-ETS. The graph suggests that two different forecasting or modeling methods, EMD-ARIMA and EMD-ETS, have been implemented to same original time series data of direct tax. Hence, the EMD-ARIMA method outperforms EMD-ETS in terms of predictive performance. After a reconstructed series, the RBFNN model is applied to forecast. The first six components and one residual are forecasted separately. Components have been recreated, and all denoised hybrid models and decomposed-hybrid models have been compared. The results reveal that denoising with wavelet analysis produced the best results, giving WA-ARIMA the lowest MSE.

The prediction performance of WA-ARIMA is the best among all models, DWT proved best for this series, along with the RBFNN model. Because the prediction performance of one-stage models is worse as compared to WA-ARIMA model, the WA-ARIMA model is best model for computing the direct tax.

### 3.4 Conclusions

Denoised and decomposed hybrid models play a pivotal role in time series analysis for several reasons. They significantly enhance the quality of the data under examination. Time series data often contain inherent noise, irregular fluctuations, and complex patterns that can obscure meaningful information. The denoising step in these hybrid models utilizes techniques such as thresholding to selectively attenuate or eliminate noisy components whereas the decomposition technique decomposes the time series into its constituent parts, such as Intrinsic Mode Functions (IMFs). In this section, we have evaluated the outcomes of the the suggested decomposed hybrid models' prediction error i.e., EMD-ARIMA, EMD-ETS, EMD-RBFNN in comparison with one-stage models and two-stage denoised hybrid models i.e., WA-ARIMA, WA-ETS, WA-RBFNN which gives us the following results shown in the table 3.4:

Table 3.4: The evaluation of the prediction error of suggested WA-ARIMA, WA-ETS, WA-RBFNN, EMD-ARIMA, EMD-ETS, EMD-RBFNN in comparison with one-stage models

Model Name	Model	MSE
ETS	one stage	8665.726
ARIMA	one stage	9612.198
RBFNN	one stage	9648.18
<b>WA-ARIMA</b>	<b>two stage</b>	<b>127.8553</b>
WA-ETS	two stage	164.4816
WA-RBFNN	two stage	2077.513
EMD-ETS	one stage	288.9586
EMD-ARIMA	two stage	184.2731
EMD-RBFNN	two stage	9647.851

In this section, numerous approaches were investigated and carried out a thorough comparison analysis in an effort to improve the accuracy of projecting direct tax receipts. The outcome of WA-ETS is 164.4816, WA-ARIMA is 127.8553 and WA-RBFNN is 2077.513 which clearly shows the WA-ARIMA gives the best prediction accuracy result in comparison to EMD-ARIMA, EMD-ETS and EMD-RBFNN. However, the EMD-ARIMA mode performs better as compared to EMD-ETS and EMD-RBFNN. This important discovery highlights the

possibility of combining AutoRegressive Integrated Moving Average (ARIMA) and Empirical Mode Decomposition (EMD) for dramatically improved forecasting accuracy. Using this methodology as a foundation, this study expanded further to improve accurate prediction by utilizing power of Radial Basis Function Neural Network (RBFNN) model. With this sophisticated strategy, the first six components and one residual of the decomposed series are projected separately. Now, in this study, reconstitute these precisely anticipated components after carefully evaluating a wide range of denoised hybrid models and decomposed hybrid models to determine their efficacy. The thorough investigation has conclusively shown that using Wavelet investigation (WA) for denoising purposes produces the most convincing results. With the lowest Mean Squared Error (MSE) of all the models taken into account, WA-ARIMA has proven to be the best at projecting direct tax income as the result of WA-ARIMA is 127.8553. Due to the skillful blending of the powerful RBFNN model with the discrete wavelet transform (DWT), WA-ARIMA has remarkable prediction performance. These findings make it clear that when compared to the powerful WA-ARIMA model, the one-stage models fall short. The WA-ARIMA model stands out as the best option in the calculation of direct tax revenues, giving fiscal authorities and policymakers the invaluable tools they need to make well-informed choices, allocate resources as efficiently as possible, and direct the trajectory of a country's finances with unwavering assurance and unmatched accuracy.

# Chapter 4

## Three stage models

### 4.1 Introduction

The decomposition of complex time series data is accomplished using an advanced signal processing technique known as Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN). It is a development of Ensemble Empirical Mode Decomposition (EEMD) and Empirical Mode Decomposition (EMD) designed to solve some of these techniques' drawbacks. CEEMDAN excels at evaluating non-stationary and non-linear time series data. The main reason for using CEEMDAN is its ability to handle adaptive noise. The results are more reliable since they help to reduce the impacts of noise and keep it from interfering with the decomposition process. By offering a thorough decomposition, CEEMDAN makes sure that all of the energy in the data is taken into account. This is essential for preserving data integrity and comprehending the underlying structure of the data. The CEEMDAN ensemble technique improves the stability of the decomposition by lowering the susceptibility to parameter selection and random fluctuations. CEEMDAN is a flexible method that can be used in a variety of fields where non-linear and non-stationary data is common i.e., finance, environmental research, biological signal processing, and more. The benefits of using CEEMDAN with denoising and decomposition techniques to ARIMA, ETS, and RBFNN models improved noise reduction, enhanced feature extraction, robustness to complex data, stability through ensemble methods, and the ability to perform multiscale analysis. The comprehensive strategy could lead to more precise and trustworthy time series forecasting and analysis across a variety of fields.

For the purpose of finding mechanical flaws in a turbofan machine in Algeria's largest fertilizer business, [Tarek et al. \(2020\)](#) utilized two different techniques, EMD and CEEMDAN. This research aimed to suggest a contrasting evaluation of three complex signal-processing techniques for the vibratory diagnostics of rotating machines working in industrial contexts. The fundamental finding of this study was that modern signal processing techniques could be easily applied to signals obtained in an industrial context and expanded to detect mechanical faults under real-world operating conditions that are more realistic than those produced on laboratory test rigs. Results from EMD and CEEMDAN were essentially equal, with CEEMDAN occasionally offering more clarity. [Aldousari et al. \(2021\)](#) suggested and compared different prediction models. EEMD and CEEMDAN techniques have been utilized to improve forecasting accuracy. Outcomes suggested that hybrid technique performed better as compared to existing one and two-stage models in predicting COVID-19 instances. A hybrid model based on complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) was used by [Aamir et al. \(2018\)](#) to forecast global crude oil prices. For this reason, the original time series of the price of crude oil was split into several short finite series known as intrinsic mode functions (IMFs). Each extracted IMF subjected to ARIMA model to estimate the parameters. All IMFs were then combined to produce the final results. The proposed method tested as well as validated using two crude oil price series: Brent and WTI (West Texas Intermediate). Two hybrid models (EMD-CEEMDAN-EBT-MM and WA-CEEMDAN-EBT-MM) flourish to improve the prediction of mineral production were cited as part of the [Qurban et al. \(2022\)](#) initiative. The data was first denoised using wavelet analysis (WA) and empirical mode decomposition (EMD). Secondly, ensemble empirical mode decomposition (EEMD) and full ensemble empirical mode decomposition (CEEMDAN) were used to decompose nonstationary data into the intrinsic mode function (IMF). The data-driven model then used the condensed noise from the noise-dominated IMFs that were subjected to empirical Bayesian threshold (EBT) application. The stochastic model then incorporates additional noise-free IMFs as input for mineral prediction. The final projection was made using all of the projected IMFs. By utilizing Pakistan's four primary mineral resources, the suggested approach is demonstrated. Mean relative error, mean square error and mean absolute percentage error were the three methods used to evaluate each model's

ability to forecast the future. With the accurate mean absolute percentage error, proposed framework WA-CEEMDAN-EBT-MM surpassed other current models in terms of mineral prediction accuracy for all four minerals. Consequently, proposed method provides a powerful mechanism for forecasting noisy and nonstationary time-series data. The development of policies and planning for the management of mineral resources would benefit from it, according to policymakers. For predicting the topological nature of gas-liquid mixtures in rectangular channels, [Yang et al. \(2023\)](#) used neural network algorithms (i.e., RBF neural network) and signal processing techniques (i.e., complete ensemble empirical mode decomposition CEEMDAN with adaptive noise-variational mode decomposition). To start, the initial topological nature time series were secondarily decomposed using full ensemble empirical mode decomposition with adaptive noise and variational mode decomposition to lessen randomness and volatility of the original signal. The signals that have been decomposed are then fed into an RBF neural network for modeling and prediction. In terms of prediction accuracy, the final findings support the suggested hybrid model's superiority over other models. [Feng et al. \(2022\)](#) developed an efficient hybrid forecasting model based on CEEMDAN to address the high nonlinear and nonstationary elements of solar output. The CEEMDAN approach separates the solar output into a number of subseries with information at various frequencies. The outcomes showed that the CEEMDAN-based model performed better on many indices than the standalone single forecasting model. Hybrid techniques could be the same, for instance, by using a uniquely built neural network that combines both linear and nonlinear models. The development and improvement of the Hybrid time series forecasting model ([Khashei and Bijari, 2012](#)) has taken a lot of work since 1968. Here, two and three-stage hybrid models are proposed using WA and decomposition methods.

[Nazar et al. \(2018\)](#) used a novel three-step hybrid intelligent prediction model that combines a set of intelligent modeling techniques and a feature extraction algorithm. In the beginning, the original data was subjected to ensemble empirical mode decomposition in order to facilitate model fitting. Then, for modeling the retrieved features, neural networks and support vector regression were proposed separately. Finally, for producing a single prediction, a weighted ensemble average employing a genetic algorithm to optimize and select the weight is proposed. The experimental results showed that suggested technique performed exceptionally well and

greatly outperformed the standard error metrics of MSE, RMSE, and MAE.

## 4.2 Decomposition Techniques

Ensemble Empirical Mode Decomposition (EEMD) and Complete Ensemble Empirical Mode Decomposition are two signal processing techniques used for examining and decomposing time-series data that are non-stationary and non-linear. These methods are particularly beneficial for data analysis applications requiring complex and noisy signals. EEMD is employed to decompose a complex time-series signal into a set of simpler oscillatory components called Intrinsic Mode Functions (IMFs). The main motive behind using EEMD is to break down a time series into its constituent parts, which are typically oscillatory and can help reveal underlying patterns, trends, or periodicities within the data. On the other hand, CEEMDAN is an extension of EEMD that aims for enhancing the decomposition of time-series data further. By boosting the decomposition's versatility and noise-handling capabilities, it tackles some of the drawbacks of EEMD. In this section, the CEEMDAN decomposition technique is used in this research to make the outcomes more accurate and enhanced.

### 4.2.1 Complete Ensemble Empirical Mode Decomposition with Adaptive Noise

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is an Empirical Mode Decomposition (EMD) technique extension that seeks to solve some of the shortcomings of classic EMD. The following are CEEMDAN's benefits and drawbacks:

#### **Advantages of CEEMDAN:**

1. **Improved Mode Separation:** CEEMDAN seeks to reduce the mode mixing issue that might arise in classical EMD. When the intrinsic mode functions (IMFs) recovered by EMD contain mixed components from distinct underlying sources, mode mixing occurs. CEEMDAN frequently allows for greater separation of these components.
2. **Adaptive Noise Enhancement:** In the decomposition process, CEEMDAN incorporates an adaptive noise component. This noise reduces the endpoint and boundary



effects associated with EMD, making the decomposition less sensitive to the beginning and finish of the data. It can also aid in mode separation.

3. **Reduced Subjectivity:** CEEMDAN frequently necessitates less subjective decision-making than classic EMD. Choosing the stopping condition for the decomposition (i.e., how many IMFs to keep) in EMD can be relatively arbitrary. CEEMDAN eliminates the requirement for manual involvement in determining the number of IMFs by including an adaptive noise component.
4. **Non-Stationary Data Applicability:** CEEMDAN is helpful for a variety of signal processing and analytical applications since it can decompose time series data that is non-stationary and nonlinear.

**Disadvantages of CEEMDAN:**

5. **Complexity of Computation:** CEEMDAN can be computationally expensive, particularly handling enormous series with large time series. The ensemble technique requires numerous iterations of the EMD process, which can be time-consuming.
6. **Variable Tuning:** CEEMDAN requires the user to configure parameters such as the ensemble size, which may require some fine-tuning to produce the best results. The selection of these parameters can have an impact on the quality of the decomposition.
7. **Sensitivity to Noise:** CEEMDAN's success is dependent on the proper estimation of the adaptive noise component. CEEMDAN's performance may degrade if the noise level is difficult to determine.
8. **Interpretation of Complexity:** It can be difficult to interpret the CEEMDAN components, including both the IMFs and the adaptive noise. Users may need to examine the results in order to extract useful information carefully.

However, there are certain drawbacks, such as increased computing complexity and parameter adjustment needs. CEEMDAN is a useful tool for dissecting complex, non-stationary time series data; nevertheless, users should be aware of its limits as well as the importance of careful parameter selection and result interpretation.

Empirical mode decomposition (EMD), an adaptive data analysis technique that deals with non-linear and non-stationary signals, was developed by [Huang et al. \(2017\)](#). The fundamental idea behind EMD was to break down a complicated signal based on its local characteristics into a small, frequently constrained set of intrinsic mode functions (IMFs). To address the problem of mode mixing in EMD, [Wu and Huang \(2009\)](#) suggested an ensemble EMD (EEMD) based noise-assisted data analysis strategy. The average of an ensemble of trials used to deconstruct noise-added signals using EMD was what is referred to as the "real IMF components" in EEMD. A white Gaussian noise signal would be added to give roughly uniform reference scale distribution, simplifying the EMD procedure also reducing EMD mode mixing. Only a few iterations later, the reconstruction error brought on by white noise might not, however, be eliminated. Reliability of reconstructed time series would have an impact on the prediction accuracy. Although increasing the number of repetitions lowers the reconstructed error, there was a significant cost to the computation. [Torres et al. \(2011\)](#) created a full EEMD with adaptive noise (CEEMDAN) to reduce reconstruction error as well as computational cost. CEEMDAN's breakdown procedure is as follows:

1. Generate several noise-added series:

$$Y^i(s) = Y(s) + p_o \omega^i(s) \quad (4.1)$$

where  $Y(s)$  represents the original signal,  $\omega^i(s)$  ( $i=1, \dots, I$ ) represents different white Gaussian noise with  $N(0, 1)$  and  $p_o$  controls the signal-to-noise ratio and also called noise coefficient.

2. Decompose each  $Y_t^i$  by using EMD to get the corresponding first modes  $IMF_1^i(s)$ . By averaging all the modes, calculate CEEMDAN first mode:

$$\overline{IMF_1^i(s)} = \frac{1}{I} \sum_{i=1}^I IMF_1^i(s) \quad (4.2)$$

3. Determine first residue  $r_1(t) = Y(s) - \overline{IMF_1^i(s)}$  and decompose the noise-added residue

$r_1(s) + p_1 E_1(\omega^i(s))$  to get the second mode:

$$\overline{IMF_1^i(s)} = \frac{1}{I} \sum_{i=1}^I E_1(r_1(s) + p_1 E_1(\omega^i(s))) \quad (4.3)$$

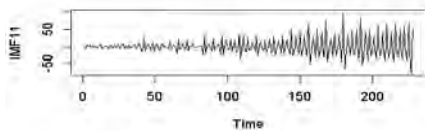
4. Repeat to obtain the remaining modes until the residue component lacks at least two extreme values

The final results demonstrate that models using the two-stage decomposition technique outperform using the CEEMDAN technique in terms of forecasting (Peng et al., 2017).

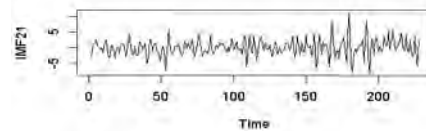
### 4.3 Results and Discussion

To obtain the local varying characteristics with respect to time from denoised time series, data of direct tax by WA and EMD are further decomposed into six components and one residual term. The CEEMDAN decomposition technique is used for eliminating further IMFs and one residual. The decomposition result of CEEMDAN of series is shown in Figure 4.1.

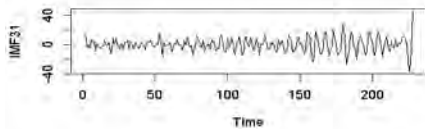
Figure 4.1: The EMD-CEEMDAN decomposition of the direct tax into five IMFs and one residual.



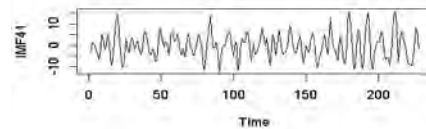
(a) IMF1 acquired by EMD-CEEMDAN



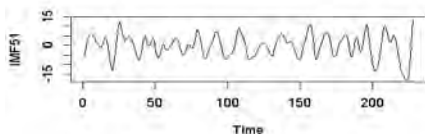
(b) IMF2 acquired by EMD-CEEMDAN



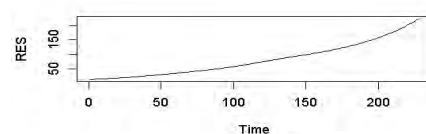
(c) IMF3 acquired by EMD-CEEMDAN



(d) IMF4 acquired by EMD-CEEMDAN



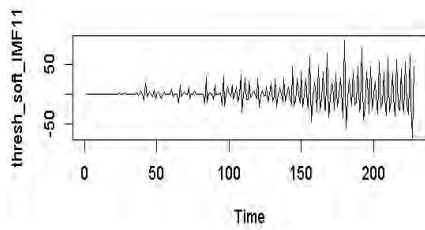
(e) IMF5 acquired by EMD-CEEMDAN



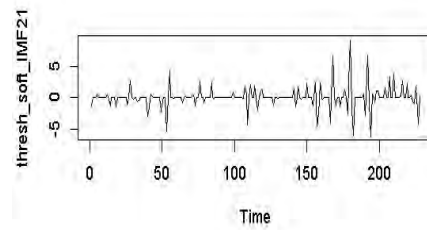
(f) IMF6 acquired by EMD-CEEMDAN

The above Fig 4.1 describes a data processing sequence involving the extraction and noise reduction of Intrinsic Mode Functions (IMFs) generated through the Empirical Mode Decomposition with Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (EMD-CEEMDAN) technique. The initial observation is that all the IMFs obtained from EMD-CEEMDAN exhibit a noisy character. This high level of noise implies that the IMFs contain significant high-frequency variations or fluctuations that could obscure the underlying patterns within the data. Therefore, a soft threshold is applied selectively to the IMFs to improve the quality of the extracted components. The process of soft thresholding is reducing or removing values that fall below a predetermined threshold while allowing data that exceed the threshold to be adjusted in a controlled manner. As a result, following the application of the threshold, it is expected that the intrinsic mode functions (IMFs) will exhibit a higher degree of smoothness and ease of handling, hence improving the capacity to identify significant patterns and trends within the data. After applying the threshold IMF the residuals are shown in Figure 4.2.

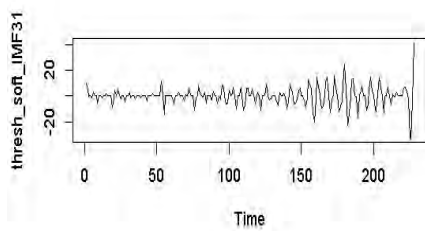
Figure 4.2: After applying the threshold, the smooth EMD-CEEMDAN decomposition of the direct tax into five IMFs and one residual.



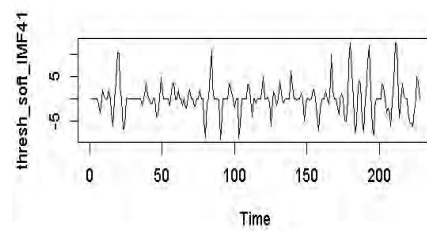
(a)Smooth IMF1 acquired by EMD-CEEMDAN



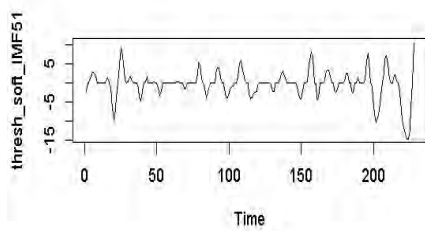
(b)Smooth IMF2 acquired by EMD-CEEMDAN



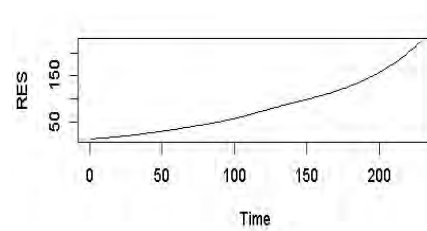
(c)Smooth IMF3 acquired by EMD-CEEMDAN



(d)Smooth IMF4 acquired by EMD-CEEMDAN



(e)Smooth IMF5 acquired by EMD-CEEMDAN

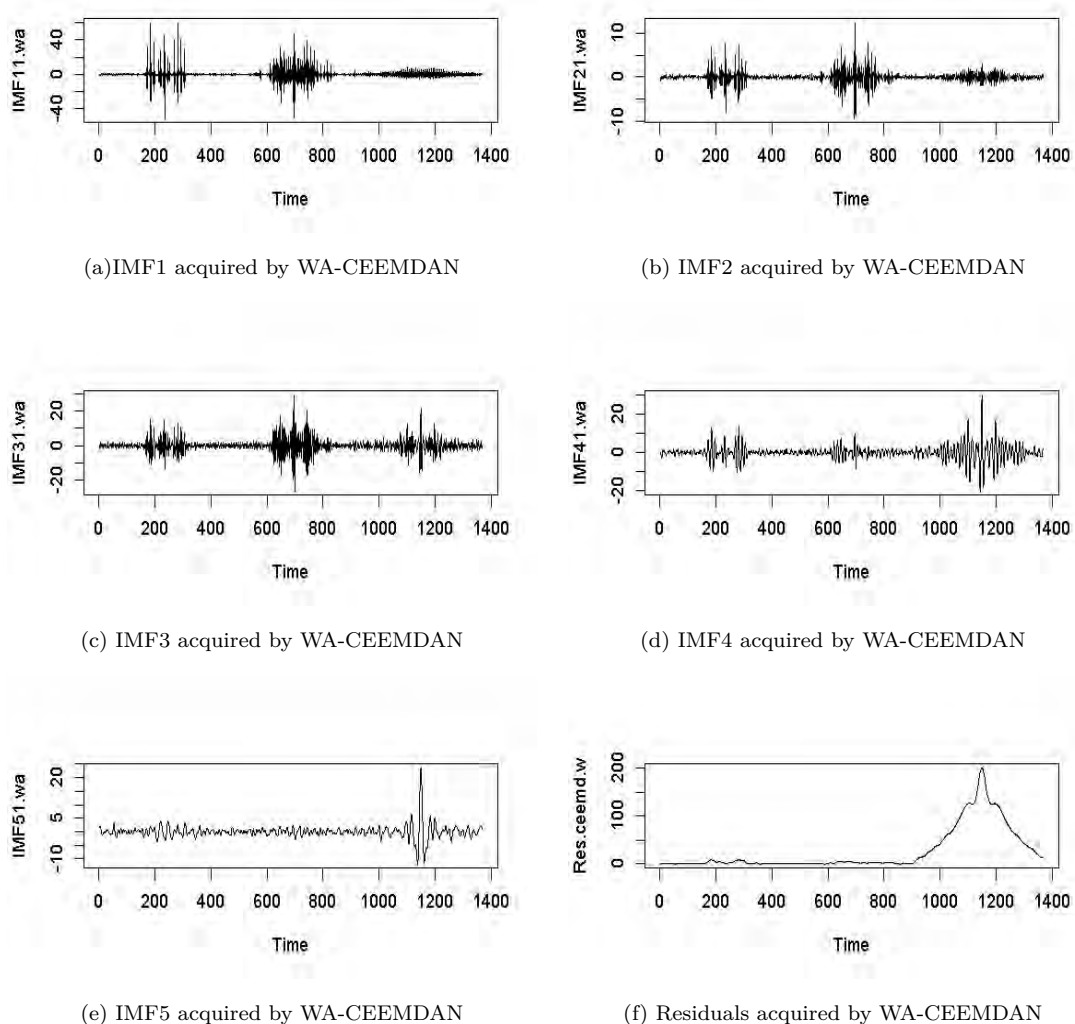


(f)Smooth IMF6 acquired by EMD-CEEMDAN

The above series is recreated when a soft threshold is applied. The soft threshold approach, often used in data preparation, involves the elimination of values below a specified threshold and the adjustment of values that are above the threshold. In Figure 4.2, the process of soft thresholding is seen to successfully mitigate the presence of high-frequency noise in the data, resulting in a smoother and simplified representation. In addition to its ability to suppress tiny or noisy oscillations in the intrinsic mode functions (IMFs), soft thresholding simplifies the representation of the data and enhances its interpretability and model ability. Soft thresholding is a technique that selectively enhances and accentuates noteworthy elements within a dataset, while simultaneously diminishing the significance of less relevant facts. This process may facilitate the extraction of features for further analysis or machine learning

activities. Therefore, the smoothed series that is obtained after applying soft thresholding might potentially be more easily understood visually and can provide a greater understanding of the underlying trends and patterns. Hence, almost all IMFs and residuals for the direct tax have different features for EMD-CEEMDAN and WA-CEEMDAN decomposition techniques. IMFs and residual for WA-CEEMDAN are shown in Figure 4.3.

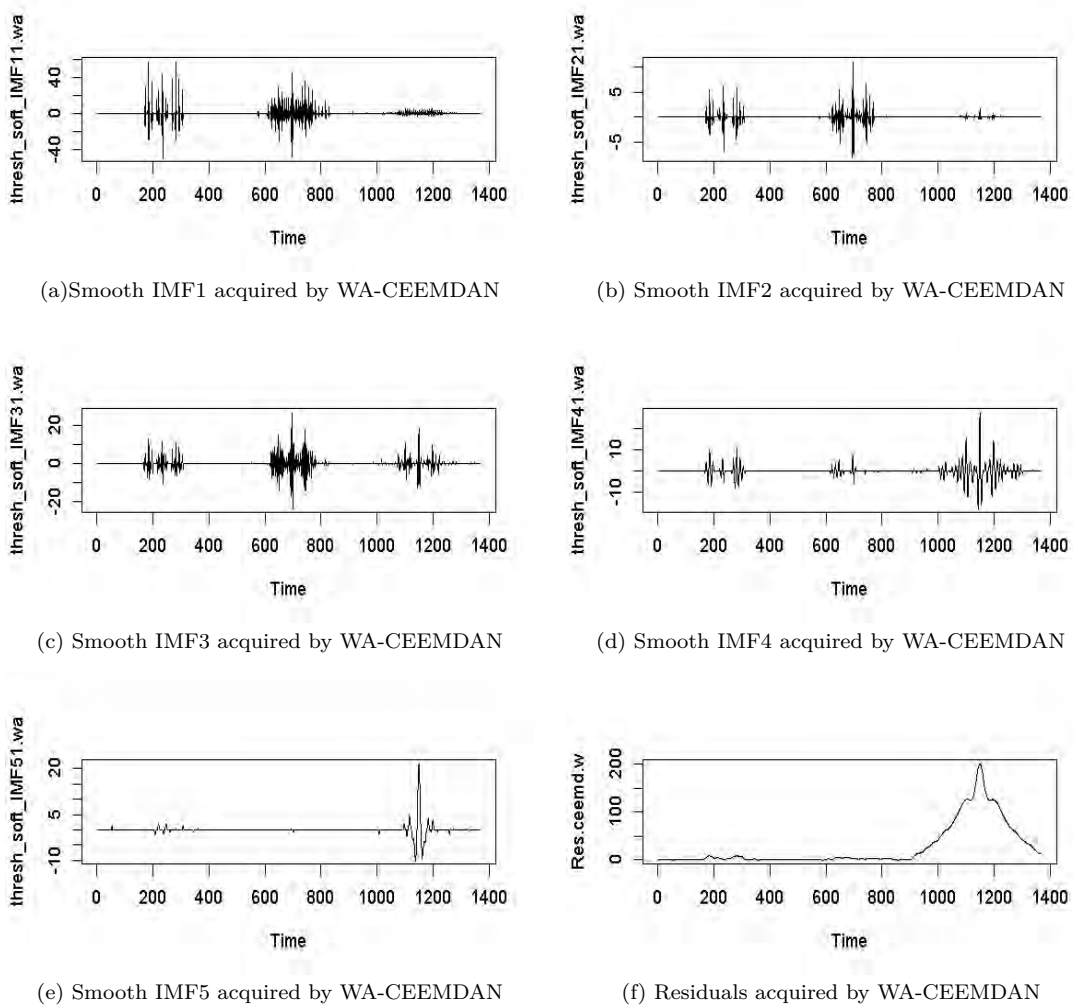
Figure 4.3: The WA-CEEMDAN decomposition of the direct tax into six IMFs and one residual



IMFs appear to exhibit high-frequency components, which often signify the presence of unwanted noise or rapid fluctuations in the data. In order to address this concern, a thresholding mechanism is implemented on these intrinsic mode functions (IMFs). The thresholding procedure is applied to the intrinsic mode functions (IMFs) in a selective manner,

resulting in the alteration or elimination of values specifically related to high-frequency oscillations. Meanwhile, the values connected with lower frequencies are either preserved or modified. The objective of this study is to mitigate or eradicate the high-frequency noise that is observed in the Intrinsic Mode Functions (IMFs), hence enhancing the quality of the data and facilitating subsequent analysis or modeling. The application of a threshold to mitigate the noise problem is expected to yield smoother and more interpretable intrinsic mode functions (IMFs). This, in turn, will enhance the accuracy and depth of analysis when examining the underlying trends and structures of the data. Smooth IMFs are shown in Figure 4.4.

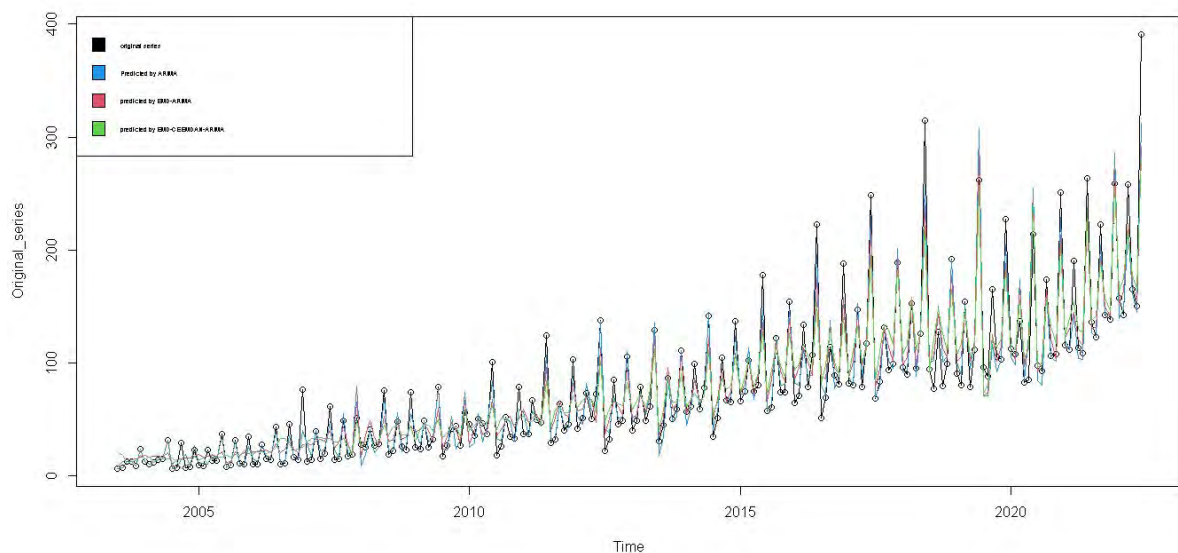
Figure 4.4: After applying the threshold, the smooth WA-CEEMDAN decomposition of the direct tax into five IMFs and one residual



By applying soft threshold to the Intrinsic Mode Functions (IMFs) that are created by a decomposition technique Wavelet Analysis with Ensemble Empirical Mode Decomposi-

tion with Adaptive Noise (WA-CEEMDAN). The primary finding indicates that all IMFs have high-frequency components, which often suggest the existence of noise or undesired fluctuations in the data. This technique entails the deliberate reduction of the amplitude of data that fall below a predetermined threshold while simultaneously changing or maintaining values that exceed the threshold in a slow and controlled way. The use of thresholding efficiently mitigates the presence of high-frequency noise inside the intrinsic mode functions (IMFs). Before applying thresholding, the time series of the direct tax is decomposed into its individual components using the Wavelet-based Complex Empirical Mode Decomposition with Adaptive Noise (WA-CEEMDAN) approach. Consequently, after the implementation of the soft threshold, the result is a "smooth WA-CEEMDAN decomposition" consisting of five Intrinsic Mode Functions (IMFs) and one residual component. The intrinsic mode functions (IMFs) are expected to exhibit a higher degree of smoothness compared to the original series as a result of the noise reduction procedure.

Figure 4.5: Prediction results of original series of direct tax in comparison with ARIMA (in blue color), EMD-ARIMA (in red color) and EMD-CEEMDAN-ARIMA (in green color)

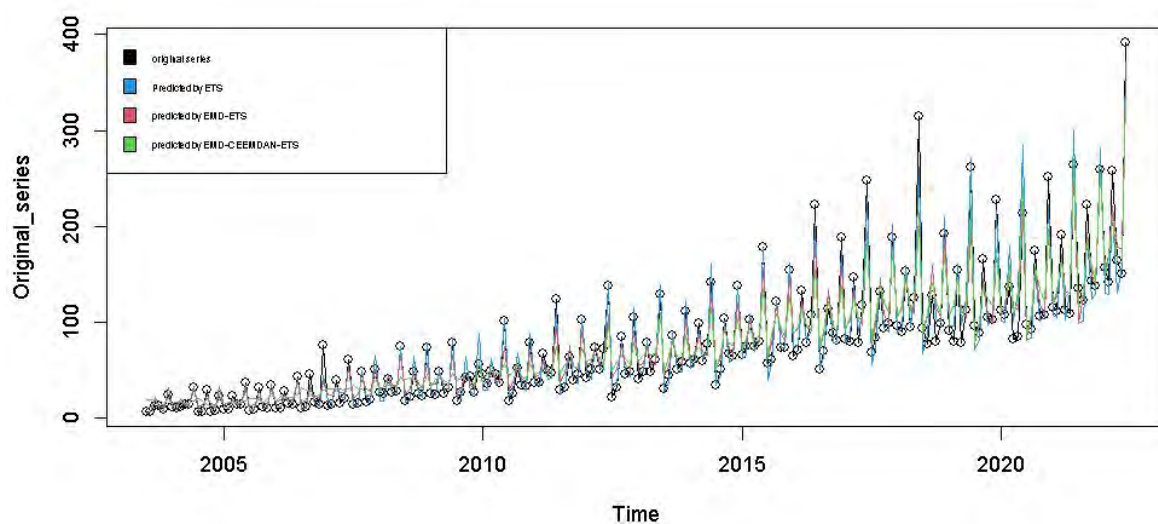


The graph illustrating the predicted outcomes of the EMD-CEEMDAN-ARIMA model is shown in Figure 4.5. This graph provides evidence supporting the superiority of the suggested EMD-CEEMDAN-ARIMA model over the ARIMA model as well as the EMD-ARIMA model. The observed enhancement can be assigned to the beneficial capabilities of the EMD-



CEEMDAN decomposition technique, which facilitates the segregation and management of intricate patterns and noise present in the data, and the subsequent utilization of ARIMA modeling, which is adept at capturing autocorrelation and trends. The hybrid technique demonstrates superior performance compared to the single ARIMA model, maybe due to its excellent handling of the intrinsic complexity and noise characteristics present in the data. Therefore, it is recommended that the EMD-CEEMDAN-ARIMA model be utilized as a more appropriate option for predicting the particular time series, as it provides enhanced precision and predictive capability.

Figure 4.6: Prediction results of original series of direct tax in comparison with ETS (in blue color), EMD-ETS (in red color), and EMD-CEEMDAN-ETS (in green color)



Similarly, the predicted graphs for EMD-CEEMDAN-ETS are shown in Figure 4.6. Figure 4.6 displays the predicted graph for EMD-CEEMDAN-ETS, indicating that the proposed EMD-CEEMDAN-ETS model outperforms ETS and EMD-ETS in comparison to the original series. The observed enhancement can be given to the combined capabilities of the EMD-CEEMDAN decomposition technique, which effectively isolates intricate patterns and noise in the data, and the subsequent utilization of ETS modeling, which is adept at capturing autocorrelation and trends. The hybrid approach demonstrates superior performance compared to the standalone ETS model, maybe due to its excellent handling of the intrinsic complexity and noise characteristics present in the data. Therefore, it is recommended that the EMD-

CEEMDAN-ETS model be utilized as a more appropriate option for forecasting the particular time series being examined, as it provides enhanced precision and predictive capability. The effectiveness of the suggested models for the direct tax with a minimum value of prediction error has been calculated. All one-stage process-based models, two-stage hybrid models, and three-stage hybrid models have been compared, as shown in Table 4.1. The accuracy performance of the proposed model WA-CEEMDAN-ARIMA is better among all models. Therefore, DWT is proved appropriate for the series along with CEEMDAN. The accuracy performance of one-stage and two-stage hybrid models is worse than the proposed model WA-CEEMDAN-ARIMA hence, WA-CEEMDAN-ARIMA is proved to be better in calculating direct taxes.

## 4.4 Conclusion

Direct taxes are a key source of government funding and have a significant impact on a nation's budgetary environment. For successful fiscal planning and administration, direct tax forecasting must be accurate. The final outcomes are given below:

Table 4.1: The evaluation of the prediction error of proposed models WA-CEEMDAN-ARIMA, WA-CEEMDAN-ETS, WA-CEEMDAN-RBFNN and EMD-CEEMDAN-ARIMA, EMD-CEEMDAN-ETS, EMD-CEEMDAN-RBFNN in comparison with all one-stage and two-stage hybrid model

Model Name	Model	MSE
ETS	one stage	8665.726
ARIMA	one stage	9612.198
RBFNN	one stage	9648.13
EMD-ARIMA	two stage	184.2731
EMD-ETS	two stage	288.9586
EMD-RBFNN	two stage	9647.851
WA-ARIMA	two stage	127.8553
WA-ETS	two stage	164.4816
WA-RBFNN	two stage	2077.513
EMD-CEEMDAN-ARIMA	three stage	149.822
EMD-CEEMDAN-ETS	three stage	466.0406
EMD-CEEMDAN-RBFNN	three stage	9648.51
<b>WA-CEEMDAN-ARIMA</b>	<b>three stage</b>	<b>94.19694</b>
WA-CEEMDAN-ETS	three stage	117.3902
WA-CEEMDAN-RBFNN	three stage	5555.723

Direct taxes are a key source of government funding and have a significant impact on a nation's budgetary environment. For successful fiscal planning and administration, direct tax forecasting must be accurate. In order to reduce the difficulties involved in analyzing time series data on total tax receipts, several data decomposition and denoising techniques have become important due to the inherently stochastic nature of direct tax data. The combination of Wavelet Analysis (WA) and AutoRegressive Integrated Moving Average (ARIMA), suitably known as WA-ARIMA, is one such effective strategy that has shown exceptional potential. Previous research projects have shown that this dynamic combination routinely outperforms alternative data processing approaches. It has become a potent tool for improving the accuracy of direct tax revenue forecasts. Three-stage hybrid models have been created, nevertheless, in an effort to estimate direct tax collections with even more accuracy and resilience. These models use a second decomposition method, called Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) or Empirical Mode Decomposition with Adaptive Noise (EMD-CEEMDAN), which combines Wavelet

Analysis and these two methods. The leading candidate among these hybrid models is WA-CEEMDAN-ARIMA, which has the lowest Mean Squared Error (MSE) of 94.19694 across all models examined. The WA-CEEMDAN-ARIMA model is now unquestionably the best option among all other models for predicting direct tax revenues as a result of this ground-breaking finding. Its unrivaled capacity to denoise and decompose intricate direct tax time series data underlines its effectiveness in delivering the government's incredibly precise revenue estimates. This enables planners and officials to make well-informed choices, allocate resources efficiently, and better control the country's financial destiny. The adoption of cutting-edge data processing methods like WA-CEEMDAN-ARIMA ushers in a new era of accuracy and foresight in fiscal management as governments attempt to navigate the ever-changing economic landscape.

# Chapter 5

## Conclusions and Recommendations

### 5.1 Conclusion

Direct taxes, which are collected directly on individuals and businesses, play an important role in stimulating economic growth and contributing to a country's total economy. Income tax is one of the most important direct taxes. Governments can create money and promote individuals to work and invest by structuring income tax systems to be progressive, with higher earnings paying higher rates. This stimulates economic activity and innovation, ultimately leading to growth. However, high direct tax rates, particularly on income and capital gains, may reduce individuals' incentives to work, save, and invest. Individuals may be discouraged from obtaining more jobs or investment possibilities if a considerable amount of their income is taxed, resulting in decreased economic activity.

The primary objective of this research project was to provide a valuable contribution to the area of tax revenue forecasting. Specifically, the study sought to examine the effectiveness of a hybrid model that integrates traditional approaches with machine learning models. The purpose was to determine if this hybrid approach might substantially improve the accuracy of direct tax revenue predictions at the Federal Board of Revenue. The primary objective of this study was to provide practical insights for enhancing direct taxation via the use of more accurate revenue predictions. The primary aim of this research was to assess the efficacy of conventional time series models, such as ARIMA and ETS, in contrast to the machine learning model, the Radial Basis Function Neural Network (RBFNN), for enhancing the predictive precision of direct tax revenues. Moreover, the primary aim of this work is to investigate

the use of Wavelet Analysis and Empirical Mode Decomposition (EMD) as preprocessing methodologies that enhance the predictive precision of direct tax income. The primary aim of this work is to investigate the use of the Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) approach in conjunction with Wavelet Analysis and EMD to get the most minimal accuracy outcomes. This study is performing modeling on direct taxes in Pakistan by using different univariate models. Decomposition and denoising techniques are used, and two and three-stage hybrid models are provided. The data contain monthly direct tax observations. From univariate ARIMA, ETS AND RBFNN models are applied which are integrated with WA and EMD. In three stages, CEEMDAN uses EMD and WA with all one-stage stage models, and three-stage hybrid models are suggested for enhancing the accuracy of direct tax data. To make IMF's smooth, soft and hard thresholds are also used. The two-stage hybrid proposed model shows better results than the single-stage model by attaining the lowest value of MSE. Similarly, three-stage hybrid models are more efficient than two-stage hybrid models. It demonstrates that both proposed models give better results. In this study, univariate models are applied. We performed ARIMA Model prediction. ARIMA model includes AR, MA, and SARIMA (Seasonal Arima). ETS models, on the other hand, are especially beneficial for analyzing and forecasting time series data, which are sequences of observations gathered at regular intervals across time. RBFNNs are simply one method for forecasting time series. The model is determined by the specific properties of the data as well as the forecasting accuracy desired. The performance of one-stage models ARIMA, ETS, and RBFNN are compared. Mean Squared Error is used as a measure of forecasting accuracy in the evaluation criterion. The ETS provides accurate predictions. Denoising and decomposition techniques are applied for two-stage hybrid models, generating IMFs. The series is reconstructed after applying IMF's smooth, soft, and hard thresholding. CEEMDAN is applied, which gives further IMF, and then the series is reconstructed. Three-stage hybrid models are suggested, i.e., WA-CEEMDAN-ARIMA, WA-CEEMDAN-ETS, WA-CEEMDAN-RBFNN, EMD-CEEMDAN-ARIMA, EMD-CEEMDAN-ETS, EMD-CEEMDAN-RBFNN where WA-CEEMDAN-ARIMA is a better hybrid model which increases prediction accuracy among all other models hence it is proved that the WA-CEEMDAN-ARIMA model is the most accurate hybrid model.

## 5.2 Recommendations

It is determined with the results of suggested models that it can give reliable predictions of direct tax revenues. Remember, tax policies can be complex, and their impact might not always be immediate. A balanced approach that takes into account both revenue needs and social considerations is crucial. It's also important to involve experts, stakeholders, and the public in discussions about tax reform. Here are a few recommendations for how we can control and minimize the taxes, mainly Direct tax:

1. The government could evaluate and change the tax brackets to make them more progressive. This means that persons with higher earnings would pay a higher percentage of their income in taxes, while those with lower incomes would pay a smaller percentage.
2. Bringing more people and businesses into the tax net can assist in more evenly spreading the tax burden. This might include better income tracking, tighter enforcement, and incentives for informal firms to register and pay taxes.
3. Encouragement of adequate record-keeping and documentation for financial transactions can result in improved tax compliance and transparency.
4. Reassessing corporation tax rates and providing incentives to enterprises that contribute to the economy may stimulate additional investment and job development.
5. Provide tax breaks to stimulate investment in specific areas, regions, or industries that match the country's economic development objectives. This has the potential to drive economic growth and employment creation.
6. Consider levying a tax on financial transactions or high-end goods. These taxes have the potential to raise income while having a minimal impact on fundamental necessities.
7. Consider modifying tax exemptions and deductions to ensure they are aimed at desired socioeconomic goals. These regulations should be reviewed and updated on a regular basis to prevent abuse and maintain fairness. Implement an increasing tax system that burdens individuals with greater incomes more heavily. This can help to spread the tax burden and increase social welfare.

8. To encourage sustainable practices and create income for environmental projects, levy taxes on ecologically hazardous activities or items. Tax digital services and transactions have grown in importance in the modern economy. This has the potential to gather income from tech behemoths and e-commerce platforms.

Some other recommendations to minimize the taxes are to investigate the idea of transitioning to a consumption-based tax system, such as VAT. This may decrease the burden of direct income taxation and encourage savings.

Also, implement a tax on unimproved land value. This has the potential to stimulate more effective land use while discouraging property speculation. User fees for specific public services or facilities might be used to generate money. This can assist in covering service costs and decrease the load on direct taxation.

In addition, running campaigns to educate citizens about the importance of paying taxes and where the tax money is utilized can encourage voluntary compliance. Tackling corruption within tax agencies can enhance revenue collection and build trust among taxpayers.

There are some suggestions for future work in prediction accuracy for direct tax revenues in Pakistan:

1. Instead of employing decomposition techniques, one can create a model using simply denoising techniques; denoising can deal with noisy series considerably better.
2. The effect of direct tax revenues on Pakistan can be observed more accurately by obtaining more detailed and accurate data on various direct tax revenues.
3. Improve data collection and reporting practices to assure consistency, accuracy, and completeness of data.
4. Explore more advanced machine learning approaches and models that are dynamic to capture complex nonlinearities and interactions in data.
5. Use strong cross-validation techniques to evaluate the predictive models' performance over multiple time periods and verify their ability to be generalized.

Estimating tax revenues accurately is difficult due to the large number of variables involved. Continuous iteration and refinement of data strategy, along with specialist knowledge, will be essential for attaining the greatest results.



# Bibliography

- Aamir, M., Shabri, A., and Ishaq, M. (2018). Crude oil price forecasting by ceemdan based hybrid model of arima and kalman filter. *Jurnal Teknologi*, 80(4):67–79.
- Adhikari, R. and Agrawal, R. K. (2013). An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*.
- Aldousari, A., Qurban, M., Hussain, I., and Al-Hajeri, M. (2021). Development of novel hybrid models for the prediction of covid-19 in kuwait. *Kuwait Journal of Science*, pages 1–6.
- Alfaki, M. M. A. and Masih, S. B. (2015). Modeling and forecasting by using time series arima models. *International Journal of Engineering Research & Technology*, 4(3):914–918.
- Ali, S., Muhmand, A., and Khuhro, F. R. (2022). Legal structuring of web based trading and tax complexities in pakistan a comparative study. *International Review of Basic and Applied Sciences (IRBAS)*, 10(2):1–27.
- Alvaredo, F., Atkinson, A. B., Piketty, T., and Saez, E. (2013). The top 1 percent an international and historical perspective. *Journal of Economic perspectives*, 27(3):3–20.
- Andellini, M., Bassanelli, E., Faggiano, F., Esposito, M. T., Marino, S., and Ritrovato, M. (2021). Forecasting hospital performances using a hybrid ets-arima algorithm. In *2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT)*, pages 42–47. IEEE.
- Anderson, O. (1977). The box-jenkins approach to time series analysis. *RAIRO-Operations Research*, 11(1):3–29.

- Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, pages 106–112. IEEE.
- Atkinson, A. B. and Bourguignon, F. (1990). The design of direct taxation and family benefits. *Journal of Public Economics*, 41(1):3–29.
- Atwood, T., Drake, M. S., Myers, J. N., and Myers, L. A. (2012). Home country tax system characteristics and corporate tax avoidance: International evidence. *The Accounting Review*, 87(6):1831–1860.
- Auriol, E. and Warlters, M. (2005). Taxation base in developing countries. *Journal of Public Economics*, 89(4):625–646.
- Azam, M. and Lukman, L. (2010). Determinants of foreign direct investment in india, indonesia and pakistan: A quantitative approach. *Journal of Managerial Sciences*, 4(1):32–41.
- Banoun, B. (2020). Wealth tax: Norway. *Wealth Tax Commission International Background Paper*, 138(1):1–13.
- Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., and Ciccozzi, M. (2020). Application of the arima model on the covid-2019 epidemic dataset. *Data in brief*, 29:105340.
- Bilquees, F. (2004). Elasticity and buoyancy of the tax system in pakistan. *The Pakistan Development Review*, 43(1):73–93.
- Chaovalit, P., Gangopadhyay, A., Karabatis, G., and Chen, Z. (2011). Discrete wavelet transform-based time series analysis and mining. *ACM Computing Surveys (CSUR)*, 43(2):1–37.
- Chatfield, S. C., Volpicelli, F. M., Adler, N. M., Kim, K. L., Jones, S. A., Francois, F., Shah, P. C., Press, R. A., and Horwitz, L. I. (2019). Bending the cost curve: time series analysis of a value transformation programme at an academic medical centre. *BMJ quality & safety*, 28(6):449–458.
- Choi, B. (2012). *ARMA model identification*. Springer Science & Business Media.

- Dietsch, P. and Rixen, T. (2015). Tax competition and global background justice. *Political theory without borders*, pages 75–106.
- Duraivel, S., Rao, A. T., Lu, C. W., Bentley, J. N., Stacey, W. C., Chestek, C. A., and Patil, P. G. (2020). Comparison of signal decomposition techniques for analysis of human cortical signals. *Journal of neural engineering*, 17(5):056014.
- Farhath, Z. A., Arputhamary, B., and Arockiam, L. (2016). A survey on arima forecasting using time series model. *Int. J. Comput. Sci. Mobile Comput*, 5(8):104–109.
- Feng, Z.-k., Huang, Q.-q., Niu, W.-j., Yang, T., Wang, J.-y., and Wen, S.-p. (2022). Multi-step-ahead solar output time series prediction with gate recurrent unit neural network using data decomposition and cooperation search algorithm. *Energy*, 261(1):125217.
- Gao, J., Nait Amar, M., Motahari, M. R., Hasanipanah, M., and Jahed Armaghani, D. (2022). Two novel combined systems for predicting the peak shear strength using rbfn and meta-heuristic computing paradigms. *Engineering with Computers*, 38(1):1–12.
- Ghauri, S. P., Ahmed, R. R., Streimikiene, D., and Streimikis, J. (2020). Forecasting exports and imports by using autoregressive (ar) with seasonal dummies and box-jenkins approaches: a case of pakistan. *Inžinerinė ekonomika*, 31(3):291–301.
- González-Audícana, M., Otazu, X., Fors, O., and Seco, A. (2005). Comparison between mallat’s and the ‘à trous’ discrete wavelet transform based algorithms for the fusion of multispectral and panchromatic images. *International Journal of Remote Sensing*, 26(3):595–614.
- Guha, B. and Bandyopadhyay, G. (2016). Gold price forecasting using arima model. *Journal of Advanced Management Science*, 4(2):11–34.
- Hills, J., De Agostini, P., and Sutherland, H. (2016). Benefits, pensions, tax credits and direct taxes. In *Social policy in a cold climate*, pages 11–34. Policy Press.
- Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the hilbert spectrum

- for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, 454(1971):903–995.
- Huang, Y., Zhu, D., Qian, Y., Zhang, Y., Porter, A. L., Liu, Y., and Guo, Y. (2017). A hybrid method to trace technology evolution pathways: a case study of 3d printing. *Scientometrics*, 111(1):185–204.
- Hussain, M. A., Siddiqui, A. A., et al. (2014). Time series forecast models for pakistan’s economic activities in past four decades. *Journal of Business Strategies*, 8(1):1.
- Jain, G. and Mallick, B. (2017). A study of time series models arima and ets. *Available at SSRN 2898968*.
- Kapeller, J., Leitch, S., and Wildauer, R. (2021). A european wealth tax for a fair and green recovery. Technical report, ICAE Working Paper Series.
- Kemal, M. A., Siddique, O., and Qasim, A. W. (2017). Fiscal consolidation and economic growth: Insights from the case of pakistan. *The Pakistan Development Review*, 56(4):349–367.
- Khashei, M. and Bijari, M. (2012). A new class of hybrid models for time series forecasting. *Expert Systems with Applications*, 39(4):4344–4357.
- Kim, D. and Oh, H.-S. (2009). Emd: A package for empirical mode decomposition and hilbert spectrum. *R J.*, 1(1):40.
- Kim, S., Maleki, N., Rezaie-Balf, M., Singh, V. P., Alizamir, M., Kim, N. W., Lee, J.-T., and Kisi, O. (2021). Assessment of the total organic carbon employing the different nature-inspired approaches in the nakdong river, south korea. *Environmental Monitoring and Assessment*, 193(7):445.
- Konarasinghe, W. (2019). Sama circular model and arima on forecasting bse sensex. *International Journal of Novel Research in Physics Chemistry & Mathematics*, 6(2):1–7.
- Krenek, A. and Schratzenstaller, M. (2018). A european net wealth tax. Technical report, WIFO Working Papers.

- López-Martín, C. (2015). Predictive accuracy comparison between neural networks and statistical regression for development effort of software projects. *Applied Soft Computing*, 27(1):434–449.
- Lu, Y. and AbouRizk, S. (2009). Automated box–jenkins forecasting modelling. *Automation in Construction*, 18(5):547–558.
- Martinez, I. (2017). Beggar-thy-neighbour tax cuts: Mobility after a local income and wealth tax reform in switzerland. *Luxembourg Institute of Socio-Economic Research (LISER) Working Paper Series*, 8.
- Moh'd, M. A., Perry, L. G., and Rimbey, J. N. (1998). The impact of ownership structure on corporate debt policy: A time-series cross-sectional analysis. *Financial Review*, 33(3):85–98.
- Molapo, M. A., Olaomi, J. O., and Ama, N. O. (2019). Bayesian vector auto-regression method as an alternative technique for forecasting south african tax revenue. *Southern African Business Review*, 23(1):2–24.
- Mondal, P., Shit, L., and Goswami, S. (2014). Study of effectiveness of time series modeling (arima) in forecasting stock prices. *International Journal of Computer Science, Engineering and Applications*, 4(2):13.
- Munir, K. and Sultan, M. (2018). Are some taxes better for growth in pakistan? a time series analysis. *International journal of social economics*, 45(10):1439–1452.
- Nazar, M. S., Fard, A. E., Heidari, A., Shafie-khah, M., and Catalão, J. P. (2018). Hybrid model using three-stage algorithm for simultaneous load and price forecasting. *Electric Power Systems Research*, 165(1):214–228.
- Nelson, C. R. and Plosser, C. R. (1982). Trends and random walks in macroeconomic time series: some evidence and implications. *Journal of monetary economics*, 10(2):139–162.
- Olkkonen, H. (2011). *Discrete Wavelet Transforms: Biomedical Applications*. BoD–Books on Demand.
- Pack, D. J. (1990). In defense of arima modeling. *International Journal of Forecasting*, 6(2):211–218.

- Panigrahi, S. and Behera, H. S. (2017). A hybrid ets–ann model for time series forecasting. *Engineering applications of artificial intelligence*, 66(1):49–59.
- Panigrahi, S., Pattanayak, R. M., Sethy, P. K., and Behera, S. K. (2021). Forecasting of sunspot time series using a hybridization of arima, ets and svm methods. *Solar Physics*, 296(1):1–19.
- Peng, T., Zhou, J., Zhang, C., and Zheng, Y. (2017). Multi-step ahead wind speed forecasting using a hybrid model based on two-stage decomposition technique and adaboost-extreme learning machine. *Energy Conversion and Management*, 153(1):589–602.
- Piketty, T. (2015). Capital, inequality and justice: Reflections on capital in the twenty-first century. *Basic Income Studies*, 10(1):141–156.
- Qasim, A. W., Kemal, M. A., Siddique, O., et al. (2015). *Fiscal consolidation and economic growth: A case study of Pakistan*. Citeseer.
- Qurban, M., Almazah, M., Nazir, H. M., Hussain, I., Ismail, M., Al-Duais, F. S., Amjad, S., Murshed, M. N., et al. (2022). Improvement towards prediction accuracy of principle mineral resources using threshold. *Mathematical Problems in Engineering*, 2022.
- Qurban, M., Zhang, X., Nazir, H. M., Hussain, I., Faisal, M., Elashkar, E. E., Khader, J. A., Soudagar, S. S., Shoukry, A. M., and Al-Deek, F. F. (2021). Development of hybrid methods for prediction of principal mineral resources. *Mathematical Problems in Engineering*, 2021:1–17.
- Rabbani, H., Mahjoob, M. P., Farahabadi, E., and Farahabadi, A. (2011). R peak detection in electrocardiogram signal based on an optimal combination of wavelet transform, hilbert transform, and adaptive thresholding. *Journal of medical signals and sensors*, 1(2):91.
- Ramos, P., Santos, N., and Rebelo, R. (2015). Performance of state space and arima models for consumer retail sales forecasting. *Robotics and computer-integrated manufacturing*, 34:151–163.
- Saad, N. (2014). Tax knowledge, tax complexity and tax compliance: Taxpayers’ view. *Procedia-Social and Behavioral Sciences*, 109(1):1069–1075.

- Salanie, B. (2011). *The economics of taxation*. MIT press.
- Sang, Y.-F. (2013). A review on the applications of wavelet transform in hydrology time series analysis. *Atmospheric research*, 122(1):8–15.
- Scott, F. R. (1955). The constitutional background of taxation agreements. *McGill LJ*, 2:1.
- Shafiq, M. N., Hua, L., Bhatti, M. A., and Gillani, S. (2021). Impact of taxation on foreign direct investment: empirical evidence from pakistan. *Pakistan Journal of Humanities and Social Sciences*, 9(1):10–18.
- Shumway, R. H., Stoffer, D. S., and Stoffer, D. S. (2000). *Time series analysis and its applications*, volume 3. Springer.
- Singh, P. and Borah, B. (2014). Forecasting stock index price based on m-factors fuzzy time series and particle swarm optimization. *International Journal of Approximate Reasoning*, 55(3):812–833.
- Soares, M. J. and Aguiar-Conraria, L. (2006). The continuous wavelet transform: A primer.
- Streimikiene, D., Rizwan Raheem, A., Vveinhardt, J., Pervaiz Ghauri, S., and Zahid, S. (2018). Forecasting tax revenues using time series techniques—a case of pakistan. *conomic research-Ekonomska istraživanja*, 31(1):722–754.
- Sun, Y., Cao, Y., Zhou, M., Wen, T., Li, P., and Roberts, C. (2019). A hybrid method for life prediction of railway relays based on multi-layer decomposition and rbfnn. *IEEE Access*, 7:44761–44770.
- Sun, Z. (2020). Comparison of trend forecast using arima and ets models for s&p500 close price. In *Proceedings of the 2020 4th International Conference on E-Business and Internet*, pages 57–60.
- Tarek, K., Abderrazek, D., Khemissi, B. M., Cherif, D. M., Lilia, C., and Nouredine, O. (2020). Comparative study between cyclostationary analysis, emd, and ceemdan for the vibratory diagnosis of rotating machines in industrial environment. *The International Journal of Advanced Manufacturing Technology*, 109(9):2747–2775.

- Tektaş, M. (2010). Weather forecasting using anfis and arima models. *Environmental Research, Engineering and Management*, 51(1):5–10.
- Torres, M. E., Colominas, M. A., Schlotthauer, G., and Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, volume 7, pages 4144–4147. IEEE.
- Wang, H., Liu, L., Dong, S., Qian, Z., and Wei, H. (2016). A novel work zone short-term vehicle-type specific traffic speed prediction model through the hybrid emd–arima framework. *Transportmetrica B: Transport Dynamics*, 4(3):159–186.
- Wang, L., Zou, H., Su, J., Li, L., and Chaudhry, S. (2013). An arima-ann hybrid model for time series forecasting. *Systems Research and Behavioral Science*, 30(3):244–259.
- Wang, X., Zhang, N., Chen, Y., and Zhang, Y. (2018). Short-term forecasting of urban rail transit ridership based on arima and wavelet decomposition. In *AIP Conference Proceedings*, volume 1967, pages 1–6. AIP Publishing.
- Werner, R. (2012). Sunspot number prediction by an autoregressive model. *Sun and Geosphere*, 7(2):75–80.
- Wu, Z. and Huang, N. E. (2009). Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in adaptive data analysis*, 1(01):1–41.
- Yang, K., Wang, Y., Li, M., Li, X., Wang, H., and Xiao, Q. (2023). Modeling topological nature of gas–liquid mixing process inside rectangular channel using rbf-nn combined with ceemdan-vmd. *Chemical Engineering Science*, 267(5):118353.
- Yun, Z., Quan, Z., Caixin, S., Shaolan, L., Yuming, L., and Yang, S. (2008). Rbf neural network and anfis-based short-term load forecasting approach in real-time price environment. *IEEE Transactions on power systems*, 23(3):853–858.
- Zucman, G. (2015). *The hidden wealth of nations: The scourge of tax havens*. University of Chicago Press.