ENHANCED MIDAS MODELS FOR MIXED-FREQUENCY TIME-SERIES FORECASTING WITH REGIME-SWITCHING

PHAKKHAPHON SAWATKAMON

มหาวิทยาลัยเทคโนโลยีสุรนารี

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy in Applied Mathematics

Suranaree University of Technology

Academic Year 2025

แบบจำลองการเลือกตัวอย่างข้อมูลความถี่ผสมเพิ่มสมรรถนะสำหรับการพยากรณ์อนุกรมเวลาที่มีความถี่ผสมร่วมกับการสลับสถานะ

นายภัคพล สวัสกมล

มหาวิทยาลัยเทคโนโลยีสุรนารี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต

สาขาวิชาคณิตศาสตร์ประยุกต์

มหาวิทยาลัยเทคโนโลยีสุรนารี

ปีการศึกษา 2568

ENHANCED MIDAS MODELS FOR MIXED-FREQUENCY TIME-SERIES FORECASTING WITH REGIME-SWITCHING

Suranaree University of Technology has approved this thesis submitted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

Thesis Examining Committee

(Asst. Prof. Dr. Prapart Pue-on)
Chairperson

(Asst. Prof. Dr. Jessada Tanthanuch)
Member (Thesis Advisor)

(Assoc. Prof. Dr. Eckart Schulz)
Member (Thesis Co-Advisor)

(Assoc Prof. Dr. Supaporn Suksern)
Member

(Asst. Prof. Dr. Tidarut Areerak)
Member

(Assoc. Prof. Dr. Yupaporn Ruksakulpiwat)
Acting Vice Rector for Academic Affairs
and Quality Assurance

(Prof. Dr. Santi Maensiri)
Acting Dean of Institute of Science

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภัคพล สวัสกมล : แบบจำลองการเลือกตัวอย่างข้อมูลความถี่ผสมเพิ่มสมรรถนะสำหรับการพยากรณ์อนุกรมเวลาที่มีความถี่ผสมร่วมกับการสลับสถานะ (ENHANCED MIDAS MODELS FOR MIXED FOR MIXED-FREQUENCY TIME-SERIES FORECASTING WITH REGIME-SWITCHING) อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร.เจษฎา ตัณฑนุช, 118 หน้า.

คำสำคัญ: แบบจำลองอนุกรมเวลา/ข้อมูลความถี่ผสม/การสลับสถานะ

วิทยานิพนธ์นี้ต้องการศึกษาแบบจำลองการเลือกตัวอย่างข้อมูลความถี่ผสมเพิ่มสมรรถนะสำหรับการพยากรณ์อนุกรมเวลาที่มีความถี่ผสมร่วมกับการสลับสถานะ โดยการศึกษาได้พิจารณาเปรียบเทียบประสิทธิภาพแบบจำลองอนุกรมเวลา ARIMA และแบบจำลองการเลือกตัวอย่างข้อมูลความถี่ผสม MIDAS จากนั้นได้พัฒนาแบบจำลองต่าง ๆ ขึ้น ได้แก่ แบบจำลอง MIDAS-AR แบบจำลอง MIDAS-AR รวมกับเงื่อนไขขีดแบ่ง (MIDAS-AR-Threshold) แบบจำลอง MIDAS-AR ร่วมกับแบบจำลองมาร์คอฟแฝง (MIDAS-AR-HMM) และแบบจำลอง MIDAS-AR ร่วมกับ Self-Organizing-Map (MIDAS-AR-SOM) แบบจำลองดังกล่าวได้นำไปทดสอบประสิทธิภาพกับข้อมูลราคาทองคำ ข้อมูลราคาน้ำมัน และข้อมูลดัชนีค่าเงินดอลลาร์อเมริกา ตั้งแต่ปี 2001-2025 ซึ่งเป็นข้อมูลที่มีความถี่ผสมและมีลักษณะผันผวนสูง รวมถึงมีการเปลี่ยนแปลงสถานะตลาด Regime-Switching การวัดประสิทธิภาพการทำนายได้ใช้เครื่องมือ 4 อย่าง ได้แก่ ค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) ค่าคลาดเคลื่อนร้อยละสัมบูรณ์เฉลี่ย (MAPE) ค่ารากที่สองของคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) และการทดสอบ Diebold–Mariano (DM-Test) โดยแบ่งข้อมูลเป็น 3 อัตราส่วนของชุดฝึกและชุดทดสอบ ได้แก่ อัตราส่วน 70:30 75:25 และ 80:20 ผลการศึกษาพบว่าแบบจำลองที่ได้พัฒนาขึ้นมาใหม่ MIDAS-AR-SOM ให้ประสิทธิภาพในการทำนายสูงที่สุด โดยเมื่อนำมาใช้กับการแบ่งข้อมูลในอัตราส่วน 70:30 โดยมีค่า MAE= 43.973 MAPE=2.133 RMSE=61.899 และผลการทดสอบ DM-Test แสดงให้เห็นว่าแบบจำลองที่พัฒนาขึ้นมีความแม่นยำในการพยากรณ์แตกต่างจากแบบจำลองเปรียบเทียบอย่างมีนัยสำคัญทางสถิติ ทั้งนี้สรุปได้ว่าแบบจำลองที่พัฒนาขึ้นมามีประสิทธิภาพในการทำนายข้อมูลที่เกี่ยวข้องกับอนุกรมเวลาที่มีความถี่ผสมและมีความผันผวนสูงมากดีกว่าแบบจำลองอื่น ๆ ที่ใช้ในการศึกษาเปรียบเทียบ

มหาวิทยาลัยเทคโนโลยีสุรนารี

สาขาวิชาคณิตศาสตร์และภูมิสารสนเทศ
ปีการศึกษา 2568

ลายมือชื่อนักศึกษา _____
ลายมือชื่ออาจารย์ที่ปรึกษา J. Tantham
ลายมือชื่ออาจารย์ที่ปรึกษาร่วม _____

PHAKKHAPHON SAWATKAMON : ENHANCED MIDAS MODELS FOR MIXED-FREQUENCY TIME-SERIES FORECASTING WITH REGIME-SWITCHING.
THESIS ADVISOR : ASST. PROF. JESSADA TANTHANUCH, Ph.D. 118 PP.

Keyword: TIME SERIES MODEL, MIXED-FREQUENCY DATA, REGIME-SWITCHING

This dissertation aims to study an enhanced Mixed Data Sampling (MIDAS) model for forecasting mixed-frequency time series with regime-switching. The study compares the forecasting performance of the ARIMA time series model and the Mixed Data Sampling (MIDAS) model. Subsequently, several models were developed, including the MIDAS-AR model, the MIDAS-AR model with threshold conditions, the MIDAS-AR model combined with a Hidden Markov Model, and the MIDAS-AR model combined with a Self-Organizing Map (MIDAS-AR-SOM). These models were evaluated using gold price data, crude oil price data, and the U.S. Dollar Index data from 2001 to 2025, which are mixed-frequency data characterized by high volatility and regime-switching market behavior. Forecasting performance was evaluated using four measures, namely Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and the Diebold–Mariano Test (DM-Test). The data were divided into training and testing sets using three ratios: 70:30, 75:25, and 80:20. The results indicate that the newly developed MIDAS-AR-SOM model provides the highest forecasting performance. When applied to the data partition ratio of 70:30, the model achieved MAE = 43.973, MAPE = 2.133, and RMSE = 61.899. In addition, the results of the DM-Test indicate that the proposed model has statistically significant differences in forecasting accuracy compared with the benchmark models. It can therefore be concluded that the developed model provides superior forecasting performance for mixed-frequency time series data with high volatility compared with the other models considered in this study.

School of Mathematical Sciences
and Geoinformatics
Academic Year 2025

Student's Signature _____

Advisor's Signature _____

Co-Advisor's Signature _____

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to everyone who has contributed to the successful completion of this dissertation.

First and foremost, I would like to thank my advisor, Asst. Prof. Dr. Jessada Tanthanuch, for his valuable guidance, deep knowledge in mathematics, and strong support in teaching both fundamental concepts and research methods. His patience, encouragement, and expertise have greatly supported my academic growth.

I would also like to express my sincere thanks to my co-advisor, Asst. Prof. Dr. Eckart Schulz, for his careful reviews, helpful feedback, and insightful suggestions that have clearly improved the quality of this research. His constructive comments have been very beneficial.

I gratefully acknowledge the financial support from the 84th Birthday Anniversary of His Majesty King Bhumibol Adulyadej Scholarship during my undergraduate studies, the Academic Excellence Scholarship during my master's degree, and the Kitti Bundit Suranaree University of Technology Scholarship during my doctoral studies.

My sincere thanks are extended to the faculty members and staff of the School of Mathematics, Suranaree University of Technology, for their continuous support throughout my studies.

Finally, I would like to thank my family and friends for their unconditional love, patience, and constant support.

Phakkhaphon Sawatkamon

มหาวิทยาลัยเทคโนโลยีสุรนารี

CONTENTS

Page

ABSTRACT IN THAI I

ABSTRACT IN ENGLISH II

CONTENTS IV

LIST OF TABLES IX

LIST OF FIGURES XII

CHAPTER

I INTRODUCTION 1

1.1 Research Objectives 4

1.2 Scope and Limitations 4

1.3 Research Procedure 5

1.4 Results 6

II LITERATURE REVIEW 7

2.1 Stationary of Time-Series Analysis 7

2.2 White Noise 10

2.3 ACF and PACF 11

2.3.1 Autocorrelation Function: ACF 11

2.3.2 Partial Autocorrelation Function: PACF 12

2.4 Linear Process 13

2.5 Development of Time-Series Models 16

2.5.1 Autoregressive Model: (AR) 16

2.5.2 Moving Average Model (MA) 21

2.5.3 ARIMA Model 24

2.5.4 Mixed Data Sampling Model: MIDAS Model 29

2.6 Regime-Switching Model 40

2.6.1 Regime Concept in Financial Markets 40

มหาวิทยาลัยเทคโนโลยีสุรนารี

CONTENTS (Continued)

Page

2.6.2 Threshold Regime-Switching Model 41
2.6.3 Hidden Markov Model (HMM) 42
2.7 Self-Organizing Map 45
2.8 Information Criteria 47
2.8.1 Akaike Information Criterion (AIC) 47
2.8.2 Bayesian Information Criterion (BIC) 47
2.9 Forecasting Performance Evaluation 48
2.9.1 Mean Absolute Error 49
2.9.2 Mean Absolute Percentage Error 49
2.9.3 Root Mean Square Error 49
2.10 Diebold–Mariano Test: DM-Test 50
2.11 Related Research 51
III **RESEARCH METHODOLOGY** 56
3.1 Phase I: Model Development 56
3.2 Phase II: Model Performance Evaluation 57
3.2.1 Research Data 57
3.2.2 Data Splitting 58
3.2.3 Program Development 58
3.2.4 Forecast Performance Evaluation 59
IV **RESULTS** 60
4.1 Overview of Research Results 60
4.2 Results of ARIMA Model Analysis 60
4.2.1 Stationarity Testing 61
4.2.2 ARIMA Model Selection Results 61
4.2.3 Residual Diagnostic Checking 61
4.2.4 Performance Evaluation Results of the ARIMA Model on the Test Set 62

มหาวิทยาลัยเทคโนโลยีสุรนารี

CONTENTS (Continued)

Page

4.3 Results of the MIDAS Model 63
4.3.1 Analysis of AIC and BIC Values for the MIDAS Model 63
4.3.2 Residual Diagnostic Checking 64
4.3.3 Performance Evaluation of the Model on the Test Dataset 64
4.3.4 The Selected MIDAS Model 66
4.4 Results of the MIDAS-AR Model 66
4.4.1 Analysis of AIC and BIC Values for the MIDAS-AR Model . . 66
4.4.2 Residual Diagnostic Checking 67
4.4.3 Performance Evaluation of the Model on the Test Dataset 67
4.4.4 The Selected MIDAS-AR Model 69
4.5 Results of the MIDAS-AR-Threshold Model 69
4.5.1 Analysis of AIC and BIC Values for the MIDAS-AR-Threshold Model 69
4.5.2 Residual Diagnostic Checking 70
4.5.3 Performance Evaluation of the Model on the Test Dataset 70
4.5.4 Regime Determination and the Selected MIDAS-AR-Threshold Model 72
4.6 Results of the MIDAS-AR-HMM Model 73
4.6.1 Regime Classification Using the Hidden Markov Model . . . 73
4.6.2 Analysis of AIC and BIC Values for the MIDAS-AR-HMM Model 75
4.6.3 Residual Diagnostic Checking 75
4.6.4 Performance Evaluation of the Model on the Test Dataset 75
4.6.5 The Selected MIDAS-AR-HMM Model 77
4.7 Results of the MIDAS-AR-HMM-SOM Model 78
4.7.1 Regime Classification Using the Self-Organizing Map and Hidden Markov Model 78

มหาวิทยาลัยเทคโนโลยีสุรนารี

CONTENTS (Continued)

Page

4.7.2 Analysis of AIC and BIC Values for the MIDAS-AR-HMM-SOM Model . . . 79
4.7.3 Residual Diagnostic Checking . . . 79
4.7.4 Performance Evaluation of the Model on the Test Dataset 80
4.7.5 Selection of the Most Appropriate MIDAS-AR-HMM-SOM Model . . . 82
4.8 Comparison of Results under Different Train:Test Ratios . . . 83
V DISCUSSION AND CONCLUSION . . . 87
5.1 Discussion . . . 87
5.1.1 Discussion of the ARIMA Model . . . 87
5.1.2 Discussion of the MIDAS Model . . . 88
5.1.3 Discussion of the MIDAS-AR Model . . . 89
5.1.4 Discussion of the MIDAS-AR-Threshold Model . . . 89
5.1.5 Discussion of the MIDAS-AR-HMM Model . . . 90
5.1.6 Discussion of the MIDAS-AR-HMM-SOM Model . . . 92
5.2 Conclusion . . . 93
REFERENCES . . . 96
APPENDICES
APPENDIX A PARAMETER SETTINGS . . . 103
A.1 ESTIMATED PARAMETERS OF THE MIDAS MODEL . 104
A.2 ESTIMATED PARAMETERS OF THE MIDAS-AR MODEL . . . 104
A.3 ESTIMATED PARAMETERS OF THE MIDAS-AR-THRESHOLD MODEL . . . 105
A.4 ESTIMATED PARAMETERS OF THE MIDAS-AR-HMM MODEL . . . 107

มหาวิทยาลัยเทคโนโลยีสุรนารี

CONTENTS (Continued)

Page

A.5 ESTIMATED PARAMETERS OF THE MIDAS-AR-HMM-SOM MODEL 108

APPENDIX B PYTHON CODE 110

CURRICULUM VITAE 118

มหาวิทยาลัยเทคโนโลยีสุรนารี

LIST OF TABLES

Table		Page
4.1	Results of the ADF stationarity test before and after first-order differencing.	61
4.2	ARIMA Model Selection Results.	61
4.3	Ljung–Box Test Results for the ARIMA Model.	62
4.4	Forecasting Performance of the ARIMA Model.	62
4.5	Parameter Estimates of the ARIMA(0,1,1) Model for the Train:Test Ratio of 80:20.	63
4.6	AIC and BIC Values of the MIDAS Model.	63
4.7	Ljung–Box Test Results for the MIDAS Model.	64
4.8	Performance Evaluation Results of the MIDAS Model under Different Train:Test Ratios.	64
4.9	AIC and BIC Values of the MIDAS-AR Model.	67
4.10	Ljung–Box Test Results for the MIDAS-AR Model.	67
4.11	Performance Evaluation Results of the MIDAS-AR Model under Different Train:Test Ratios.	67
4.12	AIC and BIC Values of the MIDAS-AR-Threshold Model.	70
4.13	Ljung–Box Test Results for the MIDAS-AR-Threshold Model.	70
4.14	Performance Evaluation Results of the MIDAS-AR-Threshold Model under Different Train:Test Ratios.	70
4.15	Threshold Value of the MIDAS-AR-Threshold Model.	72
4.16	AIC and BIC Values of the MIDAS-AR-HMM Model.	75
4.17	Ljung–Box Test Results for the MIDAS-AR-HMM Model.	75
4.18	Performance Evaluation Results of the MIDAS-AR-HMM Model under Different Train:Test Ratios.	75
4.19	AIC and BIC Values of the MIDAS-AR-HMM-SOM Model.	79

มหาวิทยาลัยเทคโนโลยีสุรนารี

LIST OF TABLES (Continued)

Table		Page
4.20	Ljung–Box Test Results for the MIDAS-AR-HMM-SOM Model.	79
4.21	Performance Evaluation Results of the MIDAS-AR-HMM-SOM Model on the Test Dataset.	80
4.22	Diebold–Mariano Test Results for $K = 4$ (Train:Test = 70:30).	84
4.23	Diebold–Mariano Test Results for $K = 8$ (Train:Test = 70:30).	84
4.24	Diebold–Mariano Test Results for $K = 12$ (Train:Test = 70:30).	84
4.25	Diebold–Mariano Test Results for $K = 4$ (Train:Test = 75:25).	85
4.26	Diebold–Mariano Test Results for $K = 8$ (Train:Test = 75:25).	85
4.27	Diebold–Mariano Test Results for $K = 12$ (Train:Test = 75:25).	85
4.28	Diebold–Mariano Test Results for $K = 4$ (Train:Test = 80:20).	86
4.29	Diebold–Mariano Test Results for $K = 8$ (Train:Test = 80:20).	86
4.30	Diebold–Mariano Test Results for $K = 12$ (Train:Test = 80:20).	86
A.1	Estimated Parameters of the MIDAS Model	104
A.2	MIDAS Lag Weights Based on the Estimated Exponential Almon Polynomial	104
A.3	Estimated Parameters of MIDAS-AR Model	104
A.4	MIDAS-AR Lag Weights Based on the Estimated Exponential Almon Polynomial	105
A.5	Estimated Parameters of the MIDAS-AR-Threshold Model for Regime 1	105
A.6	Estimated Parameters of the MIDAS-AR-Threshold Model for Regime 2	105
A.7	MIDAS-AR-Threshold Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 1	106
A.8	MIDAS-AR-Threshold Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 2	106
A.9	Estimated Parameters of the MIDAS-AR-HMM Model for Regime 1 . .	107
A.10	Estimated Parameters of the MIDAS-AR-HMM Model for Regime 2 . .	107

มหาวิทยาลัยเทคโนโลยีสุรนารี

LIST OF TABLES (Continued)

Table		Page
A.11	MIDAS-AR-HMM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 1	107
A.12	MIDAS-AR-HMM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 2	108
A.13	Estimated Parameters of the MIDAS-AR-HMM-SOM Model for Regime 1	108
A.14	Estimated Parameters of the MIDAS-AR-HMM-SOM Model for Regime 2	108
A.15	MIDAS-AR-HMM-SOM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 1	109
A.16	MIDAS-AR-HMM-SOM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 2	109

มหาวิทยาลัยเทคโนโลยีสุรนารี

LIST OF FIGURES

Figure		Page
3.1	Phase I: Model Development.	56
3.2	Phase II: Model Performance Evaluation.	59
4.1	Forecasting Performance of ARIMA Model under Different Train:Test Ratios.	62
4.2	Forecasting Performance of the MIDAS Model under Different K Values and Train:Test Ratios.	65
4.3	Forecasting Performance of the MIDAS-AR Model under Different K Values and Train:Test Ratios.	68
4.4	Forecasting Performance of the MIDAS-AR-Threshold Model under Different K Values and Train:Test Ratios.	71
4.5	Forecasting Performance of the MIDAS-AR-HMM Model under Different K Values and Train:Test Ratios.	76
4.6	Forecasting Performance of the MIDAS-AR-HMM-SOM Model under Different K Values and Train:Test Ratios.	81
4.7	Actual and Forecasted Gold Prices of the MIDAS-AR-HMM-SOM Model during the Training and Testing Periods.	83

มหาวิทยาลัยเทคโนโลยีสุรนารี

CHAPTER I

INTRODUCTION

Uncertainty is often associated with risk and danger. Being able to know or anticipate what will happen in advance helps increase our sense of security and allows us to better control situations. From a mathematical perspective, forecasting the future involves transforming complexity into a concrete form by applying mathematical principles and tools to construct mathematical models that can be used to understand and predict various phenomena. Time-series forecasting models originate from the need to predict the future based on past data. Their foundation began in the fields of statistics and finance and was later extended into computer science and artificial intelligence. The development of time-series forecasting models has a long history. Early forecasting relied on basic statistical methods, such as smoothing techniques and trend estimation. Later, statistical models became more sophisticated, employing systematic procedures for identification, estimation, and diagnostic checking, such as the ARIMA (Autoregressive Integrated Moving Average) model (Box and Jenkins, 1976). Subsequently, more advanced models were developed, including the integration with artificial intelligence techniques, resulting in highly complex models capable of capturing intricate relationships in data (Zhang, 2003; Purohit et al., 2021). Time-series forecasting is both a science and an art, combining statistics, mathematics, and computer science to enable rational and accurate insights into the future based on past information. Time-series forecasting models are highly important as they serve as tools for predicting the future and form the foundation for decision-making in many fields. These models are widely applied in economics (e.g., inflation forecasting), business (e.g., inventory management), finance (e.g., stock price prediction), medicine (e.g., disease outbreak forecasting), and agriculture (e.g., crop yield prediction).

Time-series data in economics, finance, and other fields are often collected and published at different frequencies (mixed data). For example, gross domestic product (GDP) is typically reported quarterly, while interest rates or financial market data may be

reported monthly, weekly, or even daily. This difference in frequency complicates analysis and model construction, as such data cannot be directly combined. Ignoring these differences may lead to the loss of temporal information and result in biased analysis (Foroni and Marcellino, 2013).

Mixed Data Forecasting refers to the use of data with multiple frequencies to predict future outcomes. However, it often faces limitations such as misalignment and the ragged-edge problem, where some variables are incomplete at the end of the sample period (Foroni and Marcellino, 2013). Traditional solutions involve converting data into a common frequency through aggregation or interpolation. However, such approaches may result in a loss of information and negatively affect forecasting performance. To overcome these limitations, the Mixed Data Sampling (MIDAS) model was developed, which allows the direct use of multi-frequency data without prior transformation (Ghysels, Santa-Clara, and Valkanov, 2004). This model employs weighting structures, such as Almon lag or Beta polynomial, to link variables of different frequencies. In particular, it enables the integration of high-frequency data (e.g., daily or weekly) with lower-frequency data (e.g., monthly or yearly), allowing important temporal information from high-frequency variables to enhance the forecasting accuracy of low-frequency variables. This approach has been widely accepted in economic, financial, and public policy analysis (Bahelka and de Weerd, 2024; Virk, Javed, and Awartani, 2021). Although MIDAS is a highly effective model, the assumption that parameters remain constant over time does not reflect real-world conditions. In practice, systems often experience structural changes and nonlinearity, such as economic recessions, sudden policy changes, or disease outbreaks. As a result, the traditional MIDAS model may not adequately capture changing dynamics (Ersin and Bildirici, 2023). Regime-switching models were developed to describe systems whose behavior changes across different states (regimes), such as economic expansion, recession, war, pandemics, or transitions from stable to highly volatile markets. These approaches include Threshold models (TAR/STAR), where transitions depend on threshold variables, Smooth Transition models (STAR), where transitions occur gradually, and Hidden Markov models, where regime changes follow probabilistic transitions based on a Markov process, with observed data depending on the underlying state.

In the context of financial forecasting, gold is one of the most studied assets due to its role as a safe haven and a hedge against economic and financial uncertainty, particularly during crises. Baur and McDermott (2010) showed that gold acts as a safe haven in major financial markets, especially during stock market downturns, exhibiting a significantly negative correlation with stock returns. However, gold prices are highly volatile and exhibit nonlinear behavior. Therefore, many studies focus on returns and volatility instead of price levels, as these better capture market dynamics. In particular, log returns possess desirable statistical properties, such as stationarity (Tsay, 2010). Ghysels, Santa-Clara, and Valkanov (2006) proposed using high-frequency return data to forecast low-frequency volatility and found that combining multi-frequency data significantly improves forecasting accuracy. Engle, Ghysels, and Sohn (2013) later developed the GARCH-MIDAS model, which decomposes volatility into short-term and long-term components and demonstrates that incorporating macroeconomic variables can improve forecasting performance. In the gold market, Beckmann, Berger, and Czudaj (2015) found that the relationship between gold and financial markets is not constant but varies across different market conditions, reflecting regime changes. Addressing regime changes is therefore an important issue in time-series analysis. Hamilton (1989) introduced the Markov-Switching model to capture regime changes through transition probabilities. However, this approach requires the number of regimes to be specified in advance and assumes time-invariant transitions. In highly complex and volatile markets, specifying regimes in advance can be difficult. Therefore, recent studies have increasingly applied unsupervised learning techniques to identify patterns directly from data. One such method is the Self-Organizing Map (SOM), developed by Kohonen in the 1980s and systematically described in Kohonen (2001). SOM is particularly effective in clustering high-dimensional data while preserving topological relationships, making it suitable for identifying complex market regimes. Kiviluoto (1998) applied SOM to analyze financial market structures and demonstrated its effectiveness in clustering data patterns. Similarly, Giovanis (2010) showed that using SOM to classify and identify economic regimes can improve forecasting performance, especially during financial crises. From the literature, it is evident that return and volatility play important roles in explaining financial market behavior. Meanwhile, MIDAS is an effective

tool for handling mixed-frequency data, and SOM is a powerful technique for identifying regimes without requiring predefined structures.

Therefore, this study aims to develop a forecasting model for gold prices within a mixed-frequency framework. Log return is used as the main variable to capture price dynamics, while volatility is used as an additional variable to represent market risk. Furthermore, the Self-Organizing Map (SOM) technique is applied as an unsupervised learning method to identify market regimes from empirical data. The results are then integrated with the Mixed Data Sampling (MIDAS) model, which effectively handles mixed-frequency data. This approach is expected to improve both forecasting accuracy and the ability to capture nonlinear behavior and structural changes in financial markets.

1.1 Research Objectives

1. To develop a MIDAS model with a regime-switching structure capable of forecasting mixed-frequency data, particular under condition characterized by abrupt structural changes(regime shifts).
2. To evaluate and compare the forecasting accuracy of the proposed regime-switching MIDAS model against standard benchmark models.

1.2 Scope and Limitations

This study aims to investigate time series forecasting in the context of mixed-frequency data. The analysis begins with the use of a baseline model, namely the ARIMA model, as a benchmark for comparison. The model is then extended using the MIDAS framework, which enables variables with different frequencies to be effectively integrated within a unified analytical structure.

To further improve model performance, regime-switching techniques are incorporated, including threshold models and the Hidden Markov Model (HMM), in order to capture structural changes in the data across different periods. In addition, the Self-Organizing Map (SOM), an unsupervised neural network, is employed to cluster the data and identify

latent regimes. This approach enhances the flexibility of regime classification and allows the model to better capture nonlinear patterns in the data.

In this study, monthly gold prices are used as the dependent variable, while the U.S. Dollar Index (DXY) and oil prices (Oil Price), obtained as weekly data from Yahoo Finance, are used as independent variables. The analysis is conducted using approximately 25 years of historical data. However, the selection of independent variables is subject to certain limitations, as some economic factors are not consistently available over such a long time horizon.

Model performance is evaluated using standard metrics, including RMSE, MAE, and MAPE, as well as information criteria such as AIC and BIC, together with the Ljung–Box test to assess the adequacy of the residuals. Furthermore, model comparisons are conducted using the Diebold–Mariano test to determine whether differences in forecasting performance are statistically significant.

1.3 Research Procedure

The research work proceeded as follows:

1. To study the concepts, theories, and mathematical foundations of the MIDAS model, with particular emphasis on cases where the dependent variable is of low frequency and the independent variables are of high frequency, in order to understand the model structure, lag specification, and weighting functions.

2. To study and apply regime-switching techniques, including threshold models and the Hidden Markov Model (HMM), in combination with the MIDAS model to capture structural changes in the data and improve forecasting accuracy.

3. To develop the model by integrating the Self-Organizing Map (SOM) for clustering the data and identifying latent regimes, thereby enhancing the flexibility of regime classification and enabling the model to better capture nonlinear patterns in the data.

4. To implement the model using Python for parameter estimation, forecasting, and performance evaluation of the MIDAS model combined with regime-switching techniques and SOM, allowing for systematic application to real-world data.

5. To apply the model to real data, where monthly gold prices are used as the dependent variable, and the U.S. Dollar Index (DXY) and oil prices (Oil Price), obtained as weekly data, are used as independent variables, in order to analyze mixed-frequency relationships.

6. To evaluate and compare the performance of the proposed model with baseline and alternative models using standard metrics, including RMSE, MAE, and MAPE, as well as information criteria such as AIC and BIC, together with the Ljung–Box test and the Diebold–Mariano test to assess statistically significant differences in forecasting accuracy.

1.4 Results

It is expected that the proposed model will provide more accurate forecasts than the baseline and alternative models. The effectiveness of the proposed model is evaluated using several forecasting performance measures together with the Diebold–Mariano test to determine whether the improvement in forecasting accuracy is statistically significant.

This thesis is organized as follows.

1. Chapter II reviews the theoretical background and related literature.

2. Chapter III describes the research methodology.

3. Chapter IV presents the proposed model and the experimental results.

4. Chapter V concludes the study by summarizing the main findings, discussing their implications, and providing recommendations for future research.

CHAPTER II

LITERATURE REVIEW

This chapter presents the concepts, theories, and principles related to the research, which serve as the foundation for time series model development and statistical data analysis. The chapter covers the models used in this study, including MIDAS, Regime-Switching, and Self-Organizing Map (SOM), as well as the criteria for evaluating model performance and related studies. These are used as guidelines for conducting the research and developing models with greater accuracy and reliability.

2.1 Stationary of Time-Series Analysis

In time series analysis, stationarity is a fundamental property that is important for the analysis and modeling of time series data because most statistical models, such as the Autoregressive model, Moving Average model, and ARIMA, are based on the assumption that the data are stationary. In other words, the statistical characteristics of the stochastic process do not change over time. If this assumption is not satisfied, the parameter estimation and forecasting may become unreliable (George E. P. Box and Gwilym M. Jenkins, 1976; James D. Hamilton, 1994).

Definition of Strictly Stationary:

Definition 2.1. According to Brockwell and Davis (2016), a time series $\{Y_t\}$ is said to be *strictly stationary* if

$$(Y_1, Y_2, \ldots, Y_n) \stackrel{d}{=} (Y_{1+h}, Y_{2+h}, \ldots, Y_{n+h}) \qquad \text{for all } n \geq 1,$$

where $\stackrel{d}{=}$ denotes that the joint probability distributions are identical.

Although strict stationarity is a condition that completely characterizes the proba-

bility distribution of a stochastic process, it is a restrictive condition and difficult to verify using real data. Moreover, most time series analysis methods, such as autocorrelation function analysis and parameter estimation of models, rely only on the properties of the mean and covariance. Therefore, weak stationarity is considered a sufficient and more appropriate condition in practice.

Therefore, in the construction and analysis of time series models, the assumption of weak stationarity is commonly adopted as the main framework because it can be verified from the data and is consistent with the statistical methods used for time series analysis and forecasting.

Definition of Weak Stationarity:

Definition 2.2. According to Brockwell and Davis (2016), a time series $\{Y_t\}$ is said to be *weakly stationary* if it satisfies the following conditions:

1. The mean is constant over time:

$$E(Y_t) = \mu.$$

2. The variance is constant and finite:

$$\text{Var}(Y_t) = \sigma^2 < \infty.$$

3. The covariance depends only on the time lag h and not on the time index t:

$$\text{Cov}(Y_t, Y_{t+h}) = \gamma(h).$$

If a time series is non-stationary, it affects time series analysis, particularly in linear models. The parameters of the model may become unstable because the structure of the data changes over time, causing the estimated values to be ineffective for forecasting (Hamilton, 1994). Moreover, non-constant variance may violate the assumption of white noise, leading to errors in statistical testing and confidence intervals, and may result in the

problem of spurious regression. Granger and Newbold (1974) demonstrated that analyzing relationships between non-stationary variables may produce statistically significant results even when no true relationship exists.

The stationarity of data is commonly tested using the Unit Root Test through the Augmented Dickey–Fuller (ADF) test, which was developed by David A. Dickey and Wayne A. Fuller (1979). The ADF test can be expressed in the following form:

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

where

- ΔY_t is the first difference of the data,
- α is a constant term,
- βt is the linear trend term,
- γ is the parameter used to test for the presence of a unit root,
- p is the number of lag terms added to correct the autocorrelation problem,
- ε_t is the white noise error term.

The hypotheses of the test are given by

$$H_0 : \gamma = 0,$$

$$H_1 : \gamma < 0.$$

Rejection of the null hypothesis indicates that the time series does not contain a unit root and is therefore stationary. The decision is based on the p-value associated with the test statistic. If the p-value is less than the chosen significance level (e.g., 0.05), the null hypothesis is rejected. Otherwise, there is insufficient evidence to reject the null hypothesis, suggesting that the series contains a unit root and is non-stationary. In such cases, appropriate transformations, such as differencing, should be applied before proceeding with further time series analysis.

มหาวิทยาลัยเทคโนโลยีสุรนารี

In cases where the data are non-stationary, the data can be transformed into stationary form using several methods, such as:

- Differencing, which is commonly used to eliminate trends that change over time (Box and Jenkins, 1976).
- Log Transformation, which is used to reduce non-constant variance, especially in financial data (Tsay, 2010).
- Seasonal Differencing, which is used to remove seasonal patterns in the data (Hyndman and Athanasopoulos, 2018).

Stationarity is an important condition in time series analysis because it affects the accuracy of parameter estimation and forecasting. If the data are non-stationary, it may lead to misleading conclusions. Therefore, testing stationarity using the Augmented Dickey–Fuller (ADF) test and transforming the data into stationary form are important steps before constructing time series models.

2.2 White Noise

White noise is a fundamental stochastic process in time series analysis and is commonly used as the error assumption in time series models (Box and Jenkins, 1976).

Definition of White Noise

Definition 2.3. Let $\{\varepsilon_t\}$ be a sequence of random variables. The process $\{\varepsilon_t\}$ is called *white noise* if it satisfies the following conditions:

1. The mean is zero:

$$E(\varepsilon_t) = 0.$$

2. The variance is constant:

$$\text{Var}(\varepsilon_t) = \sigma^2.$$

3. The observations are uncorrelated:

$$\text{Cov}(\varepsilon_t, \varepsilon_s) = 0, \qquad t \neq s.$$

This process is denoted by $\varepsilon_t \sim WN(0, \sigma^2)$, which means that ε_t has zero mean, constant variance, and no autocorrelation between observations at different time periods. White noise is considered a special case of a weakly stationary process and is commonly used as the error term in time series models.

Furthermore, testing whether the residuals are white noise can be performed using the Ljung–Box test, which is used to examine the independence of data across multiple lags simultaneously (Ljung and Box, 1978).

2.3 ACF and PACF

Time series analysis not only considers the stationarity of the data, but also examines the linear relationship between observations at different time periods. This relationship can be measured using the Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF), which are fundamental tools for investigating the structure of time series data and for identifying the appropriate form of time series models.

2.3.1 Autocorrelation Function: ACF

Definition 2.4. The Autocorrelation Function (ACF) at lag k, denoted by ρ_k, is a measure of the linear relationship between the current value of a time series Y_t and its lagged value at k periods. The autocorrelation coefficient takes values between -1 and 1. In statistical time series analysis for stationary time series, the ACF at lag k (Hamilton, 1994; Cryer and Chan, 2008) is defined as the ratio between the autocovariance at lag k and the variance of the time series as follows:

$$\rho_k = \frac{\gamma_k}{\gamma_0},$$

where

- ρ_k is the autocorrelation coefficient at lag k,
- γ_k is the autocovariance at lag k,
- γ_0 is the variance of the time series.

In practice, the ACF is commonly used as a guideline for identifying the order of a Moving Average model, denoted by MA(q), by considering the pattern of the autocorrelation plot. If the ACF exhibits a cut-off after lag q, it may indicate an MA(q) model. On the other hand, if the ACF gradually decays over time, it is generally characteristic of an Autoregressive model (Box et al., 2015).

For white noise processes, which have no temporal correlation, the Autocorrelation Function (ACF) is equal to zero for all non-zero lags (Brockwell and Davis, 2016), that is,

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\text{Cov}(Y_t, Y_{t-k})}{\gamma_0} = \frac{0}{\gamma_0} = 0,$$

for all $k \neq 0$ and $\rho_0 = 1$ since it represents the correlation of the series with itself.

Demonstrating that white noise has an ACF equal to zero for all non-zero lags is important because it serves as a fundamental criterion for comparing the structure of time series data. If the data behave as white noise, no temporal relationship will be observed. Conversely, if the ACF values differ significantly from zero at certain lags, it indicates that the data contain temporal dependence, which can be used in constructing time series models. In addition, this property is commonly used to assess model adequacy. The residuals of a good model should behave as white noise, meaning that the ACF of the residuals should be close to zero at all lags. If temporal correlation remains in the residuals, it indicates that the model has not fully explained the structure of the data.

2.3.2 Partial Autocorrelation Function: PACF

Definition 2.5. The Partial Autocorrelation Function (PACF) at lag k is the linear correlation between the current value of a time series Y_t and its lagged value Y_{t-k} after controlling for the effects of the intermediate lagged variables. Mathematically, the PACF at lag k can be defined as the conditional correlation as follows (Hamilton, 1994):

$$\phi_{kk} = \text{Corr}\left(Y_t, Y_{t-k} \mid Y_{t-1}, Y_{t-2}, \ldots, Y_{t-k+1}\right).$$

In addition, the PACF can also be defined through the linear regression representation of an autoregressive process as follows:

$$Y_t = \phi_{k1}Y_{t-1} + \phi_{k2}Y_{t-2} + \cdots + \phi_{kk}Y_{t-k} + \varepsilon_t,$$

where ϕ_{kk} is the coefficient representing the direct relationship between Y_t and Y_{t-k} after controlling for the effects of the other lagged variables.

The PACF plays an important role in identifying the order of an Autoregressive model. For an AR(p) model, the PACF typically exhibits a cut-off after lag p, or approaches zero when the lag is greater than p (Box et al., 2015; Tsay, 2010).

2.4 Linear Process

A Linear Process is an important concept in time series analysis and is used to describe the characteristics of a stochastic process in which the current value of the series is influenced by past shocks or disturbances through a systematic weighting structure. This concept allows the fundamental properties of the data, such as the mean, variance, and correlation structure across different time periods, to be studied more clearly. Moreover, many time series models, such as AR, MA, and ARMA models, can be represented within the framework of a linear process under certain conditions. Therefore, this concept serves as an important foundation for studying the weak stationarity properties of time series data.

Definition of Causal Linear Process

Definition 2.6. According to Brockwell and Davis (2016), the time series $\{Y_t\}$ is called a linear process if it can be represented as

$$Y_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j},$$

for all t, where $\{\varepsilon_t\}$ is a white noise process and $\{\psi_j\}$ is a sequence of constants satisfying $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$.

The process is said to be causal if $\psi_j = 0$ for all $j < 0$ in which case the representation simplifies to $Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$.

Theorem: Stationarity of a Causal Linear Process

Theorem 2.1. *According to Brockwell and Davis (2016), let $\{Y_t\}$ be a linear process defined by*

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where $\{\varepsilon_t\} \sim WN(0, \sigma^2)$. If the sequence of coefficients $\{\psi_j\}$ is absolutely summable, that is

$$\sum_{j=0}^{\infty} |\psi_j| < \infty,$$

then the process $\{Y_t\}$ is weakly stationary.

Proof

To prove that $\{Y_t\}$ is weakly stationary, three conditions must be verified: constant mean, finite and constant variance, and an autocovariance function that depends only on the lag h.

1. Constant Mean

We examine the expectation of Y_t:

$$E[Y_t] = E\left[\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}\right].$$

Since $\sum |\psi_j| < \infty$, the Fubini's Theorem (or dominated convergence) allows us to interchange the expectation and the summation:

$$E[Y_t] = \sum_{j=0}^{\infty} \psi_j E[\varepsilon_{t-j}],$$

มหาวิทยาลัยเทคโนโลยีสุรนารี

Given that for all (White Noise property):

$$E[Y_t] = \sum_{j=0}^{\infty} \psi_j(0) = 0.$$

The mean is constant at 0, which is independent of t.

2. Finite and Constant Variance

The variance of Y_t is defined as:

$$\text{Var}(Y_t) = E[Y_t^2] - (E[Y_t])^2.$$

Since $E[Y_t] = 0$:

$$\text{Var}(Y_t) = E\left[\left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}\right)^2\right] = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \psi_j \psi_k E[\varepsilon_{t-j}, \varepsilon_{t-k}].$$

By the property of white noise, $E[\varepsilon_{t-j}, \varepsilon_{t-k}] = \sigma^2$ if $j = k$, and 0 otherwise:

$$\text{Var}(Y_t) = \sigma^2 \sum_{j=0}^{\infty} \psi_j^2.$$

Since absolute summability ($\sum |\psi_j| < \infty$) implies square summability ($\sum \psi_j^2 < \infty$), the variance is finite and independent of t.

3. Autocovariance depents only lag k

Let $\text{Cov}(Y_t, Y_{t-k}) = \gamma(k)$ for any lag $k \geq 0$:

$$\text{Cov}(Y_t, Y_{t-k}) = E[Y_t, Y_{t-k}] = E\left[\left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}\right)\left(\sum_{m=0}^{\infty} \psi_m \varepsilon_{t-k-m}\right)\right].$$

Expanding the product and applying the expectation:

$$\text{Cov}(Y_t, Y_{t-k}) = \sum_{j=0}^{\infty}\sum_{m=0}^{\infty} \psi_j \psi_m E[\varepsilon_{t-j}, \varepsilon_{t-k-m}].$$

The expectation $E[\varepsilon_{t-j}, \varepsilon_{t-k-m}]$ is non-zero (equal to σ^2) only when the time indices

match: $t - j = t - k - m$, which simplifies to $j = k + m$. Substituting j:

$$\gamma(k) = \sigma^2 \sum_{m=0}^{\infty} \psi_{k+m}\psi_m.$$

The result $\gamma(k)$ is independent of the time step and depends only on the lag k.

Conclusion; Since all three conditions are satisfied, the linear process $\{Y_t\}$ is weakly stationary.

2.5 Development of Time-Series Models

The development of Time Series Models has continuously evolved in accordance with theoretical advances and practical applications. In the early stages, the models mainly focused on describing the linear relationships within the data. Later, more advanced models were developed to accommodate more complex data structures, including nonlinearity, volatility, and changes in data regimes. The models developed in each period reflect the concepts and limitations of the methodologies available at that time. In this study, the development of important models will be presented in chronological order to illustrate the trend of model development, with emphasis only on the models related to the context of this research.

2.5.1 Autoregressive Model: (AR)

The Autoregressive Model (AR) is a fundamental model for time series analysis. The concept of the model is to explain the current value of a variable using its own past values under a linear relationship. The concept of the Autoregressive Model was systematically developed during the twentieth century and was originally based on the work of George Udny Yule in 1927. An Autoregressive Model of order p, denoted by $AR(p)$, can be written as follows (Yule, 1927; Box et al., 2015):

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t,$$

where

- Y_t represents the time series value at time t,
- ϕ_0 is a constant term,
- $\phi_1, \phi_2, \phi_3, \ldots, \phi_p$ are the model parameters,
- ε_t is the random error term at time t.

In specifying the structure of the model, ε_t is commonly assumed to be a white noise process with zero mean, constant variance, and no temporal correlation, consistent with the definition of white noise presented in Section 2.2.

Selection of the Order of the AR(p) Model Determining the order of the Autoregressive model, denoted by $AR(p)$, is an important step in constructing a time series model because the value of p represents the number of lagged observations of the dependent variable used to explain the current value. If the order is specified too low, the model may not adequately capture the structure of the data. On the other hand, if the order is specified too high, the model may become unnecessarily complex and may reduce forecasting performance (Box and Jenkins, 1976).

The order of the model is commonly determined by considering the Partial Autocorrelation Function (PACF), since the PACF of an $AR(p)$ process remains significant up to lag p and exhibits a cut-off behavior afterward. Therefore, it can be used as a guideline for determining the order of the model. However, in practice, the selection is usually considered together with information criteria such as AIC or BIC in order to confirm the appropriateness of the selected model order.

Stationary of the AR(p) process In this section, we establish the conditions under which an Autoregressive process of order p, denoted as AR(p), satisfies the requirements for weak stationarity. The derivation follows the logic of transforming an AR(p) model into a causal linear process representation.

1. The AR(p) model with backshift operator

Consider the AR(p) process defined by

$$\Phi(B)Y_t = \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2),$$

where $\Phi(B)$ denotes the autoregressive polynomial in the backshift operator B, and

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p.$$

Provided that the stationarity condition is satisfied, the process can be expressed in its infinite moving-average representation as

$$Y_t = \Phi(B)^{-1}\varepsilon_t = \Psi(B)\varepsilon_t,$$

where

$$\Psi(B) = \sum_{j=0}^{\infty} \psi_j B^j.$$

2. Partial Fraction Decomposition and The Root Condition

To determine the explicit form of the coefficients $\{\psi_j\}$, we decompose the reciprocal of the autoregressive polynomial. Let $r_1, r_2, \ldots, r_p$ be the distinct roots of the characteristic equation $\Phi(B) = 0$. The operator $\Psi(B)$ can be factored as:

$$\Psi(B) = \frac{1}{\Phi(B)} = \frac{1}{\left(1 - \frac{1}{r_1}B\right)\left(1 - \frac{1}{r_2}B\right)\ldots\left(1 - \frac{1}{r_p}B\right)}.$$

By applying partial fraction decomposition, we obtain:

$$\Psi(B) = \frac{A_1}{1 - \frac{1}{r_1}B} + \frac{A_2}{1 - \frac{1}{r_2}B} + \cdots + \frac{A_p}{1 - \frac{1}{r_p}B}.$$

Using the geometric series expansion for each term, the operator becomes:

$$\Psi(B) = \sum_{i=1}^{p} A_i \left[1 + \frac{1}{r_i}B + \left(\frac{1}{r_i}B\right)^2 + \ldots\right] = \sum_{j=0}^{\infty} \psi_j B^j.$$

This expansion is convergent if and only if $|\frac{1}{r_i}| < 1$ for all $i = 1, \ldots, p$, which leads to

the fundamental root condition:

$$|r_i| > 1, \quad \forall i.$$

This implies that all roots of the characteristic polynomial must lie outside the unit circle.

3. Proof of Absolute Summability and Stationarity

From equation $Y_t = \Phi(B)^{-1}\varepsilon_t = \Psi(B)\varepsilon_t$, the process can be written as a causal linear process:

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where the coefficients are given by $\psi_j = \sum_{i=1}^{p} A_i(\frac{1}{r_i})^j$. Given the root condition $|r_i| > 1$, there exists a constant C and the smallest absolute root r_{min} such that:

$$\sum_{j=0}^{\infty} |\psi_j| \leq \sum_{j=0}^{\infty} C \left|\frac{1}{r_{min}}\right|^j < \infty.$$

Since the sequence $\{\psi_j\}$ is shown to be absolutely summable, we can directly apply the Theorem of Stationarity for Linear Processes (as detailed in Section 2.4). This confirms that the AR(p) process is weakly stationary, as it satisfies the requirement that its first and second moments remain independent of time t.

Parameter Estimation of the Model In this study, the parameters of the AR model are estimated using the Least Squares Estimation (LSE) method. The concept of this method is to determine the parameter values that minimize the sum of squared residuals (Gujarati and Porter, 2009; Hamilton, 1994).

Consider the model

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \quad t = p+1, \ldots, n.$$

Let the residual at time t be

$$\varepsilon_t = Y_t - \phi_0 - \phi_1 Y_{t-1} - \phi_2 Y_{t-2} - \cdots - \phi_p Y_{t-p}.$$

The objective function of the Least Squares method is

$$S(\phi_0, \phi_1, \ldots, \phi_p) = \sum_{t=1}^{n} \varepsilon_t^2$$

or

$$S(\phi_0, \phi_1, \ldots, \phi_p) = \sum_{t=1}^{n} (Y_t - \phi_0 - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p})^2 .$$

The estimation of parameters using the Least Squares method is the process of finding the values of $\phi_0, \phi_1, \ldots, \phi_p$ that minimize S (Draper and Smith, 1998).

In addition, the $AR(p)$ model can be expressed in matrix form as follows:

$$\mathsf{Y} = \mathsf{X}\boldsymbol{\phi} + \boldsymbol{\varepsilon},$$

where

$$\mathsf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathsf{X} = \begin{bmatrix} 1 & Y_0 & \cdots & Y_{1-p} \\ 1 & Y_1 & \cdots & Y_{2-p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & \cdots & Y_{n-p} \end{bmatrix}, \quad \boldsymbol{\phi} = \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix}.$$

When written in matrix form, the Least Squares estimator of the parameter vector $\boldsymbol{\phi}$ is

$$\hat{\boldsymbol{\phi}} = (\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\mathsf{Y}.$$

provided that $(\mathsf{X}^T\mathsf{X})^{-1}$ exists, where X^T denotes the transpose of X.

The AR(p) model has the advantage of having a simple structure and being able to explain the relationship between the current value and past values. In addition, the Autoregressive model is an important foundation of the ARIMA model (Box et al., 2015) and modern hybrid time series models. However, the limitation of the model is that it is suitable only for linear relationships and requires the assumption that the time series data are stationary. In some cases, the structure of time series data may not depend solely on past values, but may also be influenced by past error terms. This leads to the discussion of the Moving Average (MA) model in the next section.

2.5.2 Moving Average Model (MA)

The Moving Average Model (MA) is another fundamental model in time series analysis. The concept of the model is that the current value of a variable can be explained by the current random error term and past random error terms under a linear relationship. This concept was originally developed from the work of Slutsky in 1927 (Slutsky, 1927). An MA model of order q can be written as follows:

$$Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where

- Y_t represents the time series value at time t,
- θ_0 is the constant term of the model,
- $\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the model,
- ε_t is the random error term at time t.

In the Moving Average (MA) model, ε_t is commonly assumed to be a white noise process with zero mean, constant variance, and no temporal correlation. Under the concept of the MA model, the current value of Y_t is influenced by the current and past random error terms over only a limited number of q time periods. Therefore, the MA model is considered suitable for describing the short-term relationship in time series data.

Selection of the Order of the MA(q) Model Determining the order of the Moving Average model, denoted by MA(q), is an important step in constructing a time series model because the value of q represents the number of past random error terms that still influence the current value. If the order is specified too low, the model may not adequately capture the structure of the data. On the other hand, if the order is specified too high, the model may become unnecessarily complex and reduce forecasting performance.

In general, the order of the MA model is commonly determined by considering the Autocorrelation Function (ACF), since the ACF of an MA(q) process remains significant

up to lag q (Box et al., 2015; Hamilton, 1994) and exhibits a cut-off behavior afterward. That is, for $k > q$, we have $\rho_k = 0$ Therefore, in theory, the ACF plot can be used as a preliminary guideline for determining the order of the model. However, in practice, the order of the model is usually considered together with information criteria such as AIC and BIC in order to confirm the appropriateness of the model.

Stationary of the MA(q) process The Moving Average process of order q, or MA(q), is defined as:

$$Y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q} = \Theta(B)\varepsilon_t,$$

where $\{\varepsilon_t\} \sim WN(0, \sigma^2)$ and $\Theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$. Connection to Linear Processes (Section 2.4):

The MA(q) model is inherently a special case of the linear process defined in Section 2.4. By setting the coefficients $\psi_j = \theta_j$ for $j \leq q$ and $\psi_j = 0$ for $j > q$, we can observe the following:

1) Absolute Summability: Since the MA(q) process consists of a finite number of terms, the condition $\sum_{j=0}^{\infty} |\psi_j| < \infty$ is always satisfied for any finite values of θ.

2) Stationarity Guaranteed: Following the theorem established in Section 2.4, any absolutely summable linear process is weakly stationary. Consequently, every MA(q) process is inherently weakly stationary, with a constant mean $E[Y_t] = 0$ and a time-invariant variance $\text{Var}(Y_t) = \sigma^2(1 + \theta_1^2 + \cdots + \theta_q^2)$.

Conclusion: Unlike the AR(p) process, the stationarity of an MA(q) model does not depend on the roots of its characteristic polynomial. It is stationary by definition as a finite and absolutely summable linear process. (Brockwell and Davis, 2016)

Parameter Estimation of the Model For parameter estimation in the Moving Average model, denoted by MA(q), the commonly used method is Conditional Least Squares (CLS). This method operates through a recursive calculation process to determine the parameter values that minimize the Sum of Squared Errors (SSE), as follows:

1. **Initial Condition**

   Since the MA(q) model at time $t = 1$ requires past error terms prior to the observation period, namely $\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-(q-1)}$, which cannot be observed, it is common in practice to specify the initial condition by setting these initial error terms equal to zero, that is, $\varepsilon_t = 0$ for $t \leq 0$ in order to begin the calculation process.

2. **Error Calculation**

   After specifying the initial values, the model calculates the error terms ε_t sequentially from $t = 1$ to n using the previously calculated error terms according to the following equation:

$$\varepsilon_t = Y_t - \theta_0 - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \ldots - \theta_q\varepsilon_{t-q}.$$

3. **Optimization**

   The process is iterated to determine the set of parameters $(\theta_0, \theta_1, \ldots, \theta_q)$ that minimizes the objective function or the value of

$$S(\theta_0, \theta_1, \theta_2, \ldots, \theta_q) = \sum_{t=1}^{n} \varepsilon_t^2.$$

   In general, this process relies on numerical methods implemented through statistical software in order to obtain the most accurate parameter estimates (Brockwell and Davis, 2016).

The Moving Average model, denoted by MA(q), has the advantage of appropriately explaining the effects of past random error terms on the current value. The model structure reflects the short-term relationship of time series data and naturally satisfies the stationarity property for finite-order cases. However, the limitation of the model is that parameter estimation is more difficult than that of the AR model, and the selection of the model order may not always be clear for some real datasets.

In addition, the Moving Average model is an important component of the combined AR and MA model, known as the Autoregressive Moving Average (ARMA) model.

When applied to non-stationary data through differencing, it leads to the Autoregressive Integrated Moving Average (ARIMA) model, which will be discussed in the next section.

2.5.3 ARIMA Model

The Autoregressive Integrated Moving Average (ARIMA) model is one of the most widely used time series models for the analysis and forecasting of non-stationary time series data. The model was systematically developed from the concepts of George Box and Gwilym Jenkins, which were presented in the influential work *Time Series Analysis: Forecasting and Control*, first published in 1970, and later became the foundation of the approach known as the Box–Jenkins methodology (Box and Jenkins, 1976).

The main concept of the ARIMA model is to combine the structure of the Autoregressive (AR) model and the Moving Average (MA) model with the differencing process in order to transform non-stationary data into stationary form before constructing the model. Therefore, the ARIMA model is suitable for time series data that exhibit trends or changing mean levels over time.

The general form of the ARIMA model is denoted by ARIMA(p, d, q), where

- p is the order of the Autoregressive component,
- q is the order of the Moving Average component,
- d is the number of differencing operations required to make the data stationary.

The ARIMA(p, d, q) model can be written in the general form as follows:

$$\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t,$$

where

- B is the backshift operator, where $BY_t = Y_{t-1}$,
- $(1-B)^d$ is the differencing operator of order d,
- $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \ldots - \phi_p B^p$,

- $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q$,
- ε_t is the random error term at time t.

In the structure of the model, ε_t is assumed to be a white noise process with the properties described in Section 2.2.

The ARIMA model consists of three components: the Autoregressive component AR(p), the differencing component used for non-stationary data (Integrated: I), and the Moving Average component MA(q). Each component can explain the structure of the model as follows:

1. The Autoregressive component AR(p) explains the relationship between the current value of the data and its past values over p previous time periods. The model can be written as follows:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

2. The Integrated component (I) begins with testing the stationarity of the time series data using the Augmented Dickey–Fuller (ADF) Test (as described in Section 2.1). If the data are found to be non-stationary, the differencing process is applied once ($d = 1$), and the ADF Test is performed again. If the test indicates that the data are stationary, then the parameter d is equal to 1. However, if the data remain non-stationary, higher-order differencing is applied. The differencing operator of order d can be written in terms of the Backshift Operator as follows:

   Case $d = 1$:

$$(1 - B)Y_t = Y_t - Y_{t-1}$$

   Case $d = 2$:

$$(1 - B)^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$$

   The value of d represents the number of differencing operations required to remove the trend or non-stationary characteristics of the time series data in order to prepare the data for model construction in the next step.

มหาวิทยาลัยเทคโนโลยีสุรนารี

3. The Moving Average component MA(q) explains the current value of the data using past random error terms over q previous time periods. The model can be written as follows:

$$Y_t = \theta_0 + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Combining these three components results in the ARIMA(p, d, q) model, which can be written in the following form: $\phi(B)(1-B)^d Y_t = \theta(B)\varepsilon_t$

Selection of the Order of the ARIMA(p, d, q) Model In determining the order of the ARIMA(p, d, q) model, the process begins by transforming the data into stationary form through statistical testing using the ADF Test (Dickey and Fuller, 1979; Said and Dickey, 1984). After that, the orders of the parameters p and q are considered. The order of AR(p) is identified using the Partial Autocorrelation Function (PACF), as described in the section *Selection of the Order of the AR(p) Model*, while the order of MA(q) is identified using the Autocorrelation Function (ACF), as described in the section *Selection of the Order of the MA(q) Model*.

In addition, the identification of the model order is considered together with information criteria such as AIC or BIC (Akaike, 1974; Schwarz, 1978) in order to select the model with the highest efficiency and the least complexity. Finally, the model adequacy is examined to ensure that the residual terms behave as white noise.

Stationary of the ARIMA(p, d, q) process The stationarity of an ARIMA process is fundamentally evaluated through the properties of its transformed series (Hamilton, 1994; Brockwell and Davis, 2016). The conditions for stationarity are categorized based on the order of differencing (d) as follows:

1. **Transformation and Differencing Requirements**

   The requirement for achieving a stationary series depends on the initial state of the data:

   - **Case $d = 0$:** When d is equal to zero, the original series $\{Y_t\}$ is assumed to be inherently stationary. In this scenario, no differencing transformation is

required to stabilize the statistical properties of the series, such as its mean and variance.

- **Case $d > 0$ (The ARIMA Process):** For any $d > 0$, the original series $\{Y_t\}$ is inherently non-stationary, typically characterized by the presence of one or more unit roots (stochastic trends). The application of the differencing operator $(1 - B)^d$ is necessary to eliminate these trends (Box et al., 2015; Hyndman and Athanasopoulos, 2018). While the raw series $\{Y_t\}$ lacks weak stationarity, the transformed series

$$W_t = (1 - B)^d Y_t$$

is rendered stationary.

2. **Autoregressive Stability (The Unit Root Condition)**

For the process to maintain weak stationarity (either the original series when $d = 0$, or the differenced series when $d > 0$), the characteristic equation associated with the autoregressive polynomial:

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p = 0$$

must satisfy the stability condition. Specifically, all complex roots of $\Phi(B) = 0$ must lie outside the unit circle in the complex plane (Hamilton, 1994; Brockwell and Davis, 2016). This ensures that the influence of historical observations dissipates over time rather than persisting indefinitely.

3. **Invertibility of the Moving Average Component**

Although the moving average polynomial $\theta(B)$ does not dictate stationarity, it governs the invertibility of the model. For the model to be uniquely identifiable and representable as an infinite AR process, the roots of the characteristic equation:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q = 0$$

are conventionally required to lie outside the unit circle (Box et al., 2015; Tsay, 2010). This condition ensures that the current observation can be expressed as a convergent linear combination of past observations.

Parameter Estimation of the Model After determining the appropriate order of the ARIMA(p, d, q) model, the next step is to estimate the parameters of the autoregressive and moving average components, namely $\phi_1, \phi_2, \ldots, \phi_p$ and $\theta_1, \theta_2, \ldots, \theta_q$, in order for the model to appropriately describe the characteristics of the data. In this study, the parameter estimation method used is the Least Squares Estimation method (Brockwell and Davis, 2016), in which the principle is to select the parameter values that minimize the sum of squared residuals.

For parameter estimation, the residual of the model at time t is defined as

$$e_t(\phi, \theta) = \phi(B)(1 - B)^d Y_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q},$$

which is the residual calculated from the observed data under the specified parameter values. In general, the residuals can be recursively computed from the ARIMA model equation. The objective function is then defined as the sum of squared residuals as follows:

$$S(\phi, \theta) = \sum_{t=1}^{n} e_t^2(\phi, \theta),$$

where $\phi = (\phi_1, \phi_2, \ldots, \phi_p)$ and $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$

Therefore, parameter estimation using the Least Squares method is the process of finding the estimates $\hat{\phi}$ and $\hat{\theta}$ that minimize $S(\phi, \theta)$, that is,

$$(\hat{\phi}, \hat{\theta}) = \arg\min_{\phi, \theta} S(\phi, \theta).$$

Since the parameters in the Moving Average (MA) component appear in a nonlinear form through the lagged residuals, the parameter estimates cannot be directly obtained in closed form as in ordinary linear regression. Therefore, iterative numerical optimization methods (Box et al., 2015; Nocedal and Wright, 2006) are required to search for the

optimal values, such as the Gauss–Newton method or the Levenberg–Marquardt method. The calculation process begins with initial parameter values and repeatedly updates them until the change in the objective function becomes very small or converges to a minimum point.

The ARIMA model is one of the most widely used tools in time series analysis because it has a clear structure and is supported by a strong statistical theoretical foundation. One of the main strengths of the model is its ability to handle non-stationary data through the differencing process, making it suitable for data with trends or changing mean levels over time. In addition, the model can effectively explain temporal relationships through the autoregressive (AR) and moving average (MA) components, together with a systematic estimation and diagnostic procedure based on the Box–Jenkins methodology. As a result, ARIMA is considered an effective tool for short- to medium-term forecasting of univariate time series data. However, the ARIMA model also has several important limitations. First, the model was developed under the framework of single-frequency data, which requires all variables to share the same time frequency context, such as monthly or quarterly data. Consequently, the model cannot directly integrate variables with different frequencies. Second, the model focuses on explaining linear relationships and therefore may not adequately capture complex dynamics or nonlinear relationships. Finally, increasing the lag order may lead to the problem of overparameterization, which can reduce out-of-sample forecasting performance. Particularly in economic and financial research, the data often consist of mixed-frequency variables, such as a monthly dependent variable combined with daily or quarterly explanatory variables. The ARIMA model cannot efficiently accommodate this type of data structure. For this reason, models capable of directly integrating mixed-frequency data were developed, namely the Mixed Data Sampling (MIDAS) model, which has become increasingly popular in recent years.

2.5.4 Mixed Data Sampling Model: MIDAS Model

In general econometric time series analysis, such as the ARIMA (Autoregressive Integrated Moving Average) model, there is usually a basic requirement that all variables in the system must have the same data frequency. For example, if the objective is to

forecast Gross Domestic Product (GDP), which is quarterly data, the explanatory variables must also be in quarterly frequency. However, in practice, economic and market data are often collected at different frequencies, such as monthly interest rates or daily stock price indices. If these variables are directly analyzed together with quarterly variables, a frequency mismatch problem will occur. In the past, this problem was commonly addressed through temporal aggregation, which involves converting high-frequency data into lower-frequency data (for example, averaging daily data to obtain monthly values). However, this approach often leads to the loss of important information, particularly short-term volatility and leading economic signals, thereby reducing forecasting performance (Marcellino and Schumacher, 2010). To overcome these limitations, Eric Ghysels and colleagues (2004) introduced the MIDAS (Mixed Data Sampling) model, which is a framework that allows explanatory variables with different frequencies to be incorporated into the same equation without converting them to lower-frequency data beforehand. As a result, the characteristics and information contained in the original data can be preserved more completely (Ghysels, Santa-Clara, and Valkanov, 2004).

The MIDAS model is a model that explains the relationship between a low-frequency dependent variable (Y_t) and a high-frequency independent variable (X_t) through a distributed lag weighting structure, as shown in the following equation:

$$Y_t = \beta_0 + \beta_1 \sum_{k=1}^{K} w_k(\nu) X_{t-k/m} + \varepsilon_t,$$

where

- Y_t is the dependent variable observed at low frequency at time t,
- $X_{t-k/m}$ is the independent variable observed at high frequency,
- K is the maximum number of lags considered,
- $w_k(\nu)$ is the lag weight at lag k, which depends on the parameter ν,
- β_0 is the constant term of the model,
- β_1 is the coefficient of the weighted sum,

- ε_t is the random error term, and
- m is the frequency ratio between the explanatory variable X and the dependent variable Y.

An important advantage of the MIDAS model is its ability to directly utilize the potential of high-frequency data to explain or forecast low-frequency variables, such as using daily raw data to estimate monthly economic indicators. For this reason, MIDAS has become widely used in macroeconomic, financial, and asset price forecasting research (Foroni, Marcellino, and Schumacher, 2015), and is considered an important tool in modern time series analysis that effectively reduces the limitations of single-frequency models.

Multivariate MIDAS: In real economic and social conditions, the dependent variable is often influenced simultaneously by various surrounding factors. An analysis using only a single explanatory variable may not comprehensively explain the complex relationship structure. For example, forecasting asset prices may require the simultaneous consideration of financial market indices, exchange rates, and commodity prices. For this reason, the MIDAS model was extended to the Multivariate MIDAS Regression framework in order to support the analysis involving p explanatory variables, as shown in the following equation:

$$Y_t = \beta_0 + \sum_{j=1}^{p} \beta_j \sum_{k=1}^{K} w_{j,k}(\nu_j) X_{j,t-k/m} + \varepsilon_t,$$

where

- Y_t is the dependent variable observed at low frequency at time t,
- p is the total number of explanatory variables,
- $X_{j,t-k/m}$ is the jth explanatory variable,
- K is the maximum lag considered for the jth variable,
- $w_{j,k}(\nu_j)$ is the lag weight at lag k for the jth variable, which depends on the parameter ν_j,

- β_0 is the constant term of the model,
- β_j is the coefficient of the jth explanatory variable,
- ε_t is the random error term.

From the above equation, it can be seen that each explanatory variable can have a different number of lags and weighting structure, providing the model with high flexibility in reflecting temporal effects. For example, some variables may affect Y_t in the short term, while others may exert cumulative long-term effects. MIDAS can appropriately represent these effects through the weighting function assigned to each variable. An important advantage of the multivariate MIDAS model is its ability to combine data from multiple sources with different frequencies into a single equation without the need to transform all variables into the same frequency beforehand. Consequently, this approach helps reduce information loss and improve forecasting performance, particularly when the dependent variable is influenced simultaneously by multiple factors (Foroni et al., 2015). However, as the number of explanatory variables increases, problems such as multicollinearity may arise, parameter estimation becomes more complex, and the model may be more susceptible to overfitting. Therefore, many studies apply variable selection methods, dimensionality reduction techniques, or combine MIDAS with machine learning methods in order to improve model performance. At present, the multivariate MIDAS model has become highly popular in macroeconomics, financial market forecasting, asset pricing, energy analysis, and public health data analysis because it can efficiently utilize information from multiple sources and multiple frequencies simultaneously.

MIDAS-AR: The MIDAS-AR model is an extension of the Mixed Data Sampling (MIDAS) model by incorporating autoregressive terms of the dependent variable into the equation. This allows the model to simultaneously reflect both the influence of high-frequency explanatory variables and the temporal persistence of the dependent variable. This type of model is suitable for time series data in which the current value is related to its past values, such as gross domestic product, inflation rates, asset prices, and other economic and financial variables. Clements and Galvão (2008) applied the MIDAS framework to

forecast economic growth in the United States by incorporating lagged values of the dependent variable into the forecasting equation in order to capture the persistence characteristics of macroeconomic data. This approach is considered one of the important foundations of the MIDAS-AR model and has become widely used in econometric studies. The general form of the MIDAS-AR model can be written as follows:

$$Y_t = \beta_0 + \sum_{r=1}^{P} \phi_r Y_{t-r} + \sum_{j=1}^{p} \beta_j \sum_{k=1}^{K} w_{j,k}(\nu_j) X_{j,t-k/m} + \varepsilon_t$$

where

- Y_t is the dependent variable observed at low frequency at time t,
- p is the total number of explanatory variables,
- P is the autoregressive order,
- ϕ_r is the coefficient of the rth lagged dependent variable,
- $X_{j,t-k/m}$ is the jth explanatory variable,
- K is the maximum lag considered for the jth variable,
- $w_{j,k}(\nu_j)$ is the lag weight at lag k for the jth variable, which depends on the parameter ν_j,
- β_0 is the constant term of the model,
- β_j is the coefficient of the jth explanatory variable,
- ε_t is the random error term.

An important advantage of the MIDAS-AR model is its ability to efficiently utilize mixed-frequency data while simultaneously capturing the temporal relationship of the dependent variable through its past values. Therefore, the model is suitable for data with high persistence. In addition, the model reduces the information loss that commonly occurs when high-frequency data are aggregated into lower-frequency data before analysis (Andreou, Ghysels, and Kourtellos, 2010). However, the application of the model still has important limitations, particularly the complexity of parameter estimation because the

มหาวิทยาลัยเทคโนโลยีสุรนารี

weighting functions are generally nonlinear. Consequently, estimation typically requires the use of Nonlinear Least Squares (NLS). Furthermore, caution must be taken regarding overfitting problems resulting from specifying too many lag terms, as well as the common factor restriction assumption, which may affect forecasting accuracy when the dynamic structures of the dependent and explanatory variables differ substantially. Therefore, the MIDAS-AR model is considered an appropriate approach for forecasting time series data that require both mixed-frequency explanatory variables and lagged dependent variables. The model effectively combines the advantages of autoregressive models and MIDAS models into a unified framework.

Weight Functions: An important core component of the MIDAS model is the weighting function, which determines the level of influence of high-frequency explanatory variables at different lag values on the low-frequency dependent variable. The use of weighting functions allows a large number of parameters to be estimated using a small parameter vector (ν), which helps reduce the dimensionality curse problem and prevents overfitting when the number of lag terms (K) is large (Ghysels, Sinko, and Valkanov, 2007).

In general, the weighted component of the high-frequency variable can be written in the following general form:

$$Z_t = \sum_{k=1}^{K} w_k(\nu) X_{t-k/m},$$

where

- Z_t is the weighted sum of the high-frequency explanatory variable,
- $X_{t-k/m}$ is the value of the explanatory variable at lag k (where m is the frequency ratio between the high-frequency and low-frequency data),
- $w_k(\nu)$ is the weighting function at lag k,
- ν is the parameter vector controlling the characteristics and shape of the weighting distribution.

In this study, the Exponential Almon Lag weighting function is selected because it has high flexibility and requires only two parameters for estimation, namely ν_1 and ν_2. The

weighting function can be written as follows:

$$w_k(\nu_1, \nu_2) = \frac{\exp(\nu_1 k + \nu_2 k^2)}{\sum_{j=1}^{K} \exp(\nu_1 j + \nu_2 j^2)}, \qquad k = 1, \ldots, K.$$

An important characteristic of this function is its ability to reflect various temporal effect structures, such as geometric decay or hump-shaped effects in which the influence initially increases before gradually decreasing. These characteristics are consistent with the behavior of economic and financial market variables.

In order for the estimation to have statistical meaning and to be interpreted as relative weights, the weighting function in the MIDAS model must satisfy the following fundamental theoretical properties.

Theorem: Normalization

Theorem 2.2. *Let the weighting function be defined as*

$$w_k(\nu) = \frac{g_k(\nu)}{\sum_{j=1}^{K} g_j(\nu)},$$

where $g_k(\nu) > 0$ *for all values of* k*. Then, the sum of all weights is always equal to one:*

$$\sum_{k=1}^{K} w_k(\nu) = 1,$$

where $g_k(\nu) = \exp(\nu_1 k + \nu_2 k^2)$ *and* $g_j(\nu) = \exp(\nu_1 j + \nu_2 j^2)$

Proof

Consider the sum of the weighting functions:

$$\sum_{k=1}^{K} w_k(\nu) = \sum_{k=1}^{K} \left[\frac{g_k(\nu)}{\sum_{j=1}^{K} g_j(\nu)} \right].$$

Since the denominator

$$\sum_{j=1}^{K} g_j(\nu)$$

is constant with respect to the summation index k, it can be factored out of the summation:

มหาวิทยาลัยเทคโนโลยีสุรนารี

$$\sum_{k=1}^{K} w_k(\nu) = \frac{1}{\sum_{j=1}^{K} g_j(\nu)} \sum_{k=1}^{K} g_k(\nu) = 1$$

Theorem: Positivity

Theorem 2.3. *For the Exponential Almon lag weighting function, the weights at all lag periods are always positive:*

$$w_k(\nu_1, \nu_2) > 0, \quad \forall k \in \{1, 2, \ldots, K\}.$$

Proof

From the property of the exponential function, whose range is always positive real numbers ($\exp(x) > 0$ for all $x \in \mathbb{R}$), both the numerator $\exp(\nu_1 k + \nu_2 k^2)$ and every term in the denominator are positive. Therefore, the ratio of positive real numbers is always positive.

Theorem: Boundedness

Theorem 2.4. *The weighted sum Z_t always lies within the bounds of the original data if $w_k(\nu) \geq 0$ and $\sum w_k(\nu) = 1$. That is, $\min_{k=1,\ldots,K}(X_{t-k/m}) \leq Z_t \leq \max_{k=1,\ldots,K}(X_{t-k/m})$.*

Proof

Let $m = \min(X_{t-k/m})$ and $M = \max(X_{t-k/m})$. Then, $m \leq X_{t-k/m} \leq M$. Multiplying by $w_k(\nu) \geq 0$ and summing over k gives

$$m \sum_{k=1}^{K} w_k(\nu) \leq \sum_{k=1}^{K} w_k(\nu) X_{t-k/m} \leq M \sum_{k=1}^{K} w_k(\nu).$$

Since $\sum_{k=1}^{K} w_k(\nu) = 1$, it follows that, $m \leq Z_t \leq M$.

From the mathematical properties of the weighting functions presented above, it can be seen that the specification of weights in the MIDAS model plays an important role in the performance of the model in several aspects.

First, regarding model stability, specifying that each weight is nonnegative and that the sum of all weights is equal to one causes the weighted sum Z_t to behave as a weighted average of past observations. As a result, the obtained values do not deviate

excessively from the range of the original data and help reduce fluctuations that may arise from inappropriate weighting schemes. This property is important for parameter estimation and forecasting because it contributes to greater model stability.

Second, regarding the interpretation of results, expressing Z_t as a convex combination allows a clear interpretation that it is obtained from combining past high-frequency observations with different weights assigned to different time periods. In other words, observations from some periods may exert greater influence on the dependent variable than others, depending on the form of the weighting function selected. Consequently, this helps explain the temporal impact structure of past information in a systematic manner.

In addition, the use of weighting functions helps reduce the number of parameters that need to be estimated because it is unnecessary to estimate the coefficient of each lag separately. Instead, the effects of many lag terms can be represented using only a small number of parameters. As a result, the model becomes more parsimonious and more suitable for practical applications with real-world data.

Therefore, the weighting function is not merely a technical component of the MIDAS model, but also an important mechanism that provides mathematical appropriateness, estimation stability, and clear interpretability of the model results.

Parameter Estimation of the Model The parameter estimation of the MIDAS model aims to obtain the most appropriate parameter estimates in order for the model to explain the data and provide accurate forecasting. In general, the parameters to be estimated consist of the constant term of the model, the coefficients of the explanatory variables, and the parameters of the weighting function (Ghysels et al., 2007; Andreou et al., 2010), and in some cases may also include the coefficients of lagged dependent variables. Since the parameters of the weighting function appear in a nonlinear form, they cannot be estimated directly using the Ordinary Least Squares (OLS) method. Therefore, this study uses the Nonlinear Least Squares (NLS) method (Ghysels et al., 2007; Hamilton, 1994) for parameter estimation by finding the set of parameters that minimizes the sum of squared residuals, as follows:

$$\hat{\psi} = \arg\min_{\psi} \sum_{t=1}^{T} \left(Y_t - \hat{Y}_t(\psi) \right)^2,$$

where ψ represents the vector of all model parameters, and $\hat{Y}_t(\psi)$ is the estimated value obtained from the model under that parameter set. In the numerical computation, the researcher uses the L-BFGS-B algorithm (Byrd et al., 1995; Zhu et al., 1997), which is an optimization method suitable for nonlinear objective function minimization problems and allows parameter bounds to be specified. The estimation process is performed iteratively until the objective function converges to the minimum value according to the specified criterion.

Since the MIDAS model can be classified into several forms according to the number of explanatory variables and dynamic components, this study divides the presentation of parameter estimation into three cases: the univariate MIDAS model, the multivariate MIDAS model, and the MIDAS model with autoregressive terms, as described in the following sections.

Case 1: BASIC MIDAS

The model contains only one explanatory variable (Ghysels et al., 2004). The equation can be written as follows:

$$Y_t = \beta_0 + \beta_1 \sum_{k=1}^{K} w_k(\nu) X_{t-k/m} + \varepsilon_t,$$

where

- β_0 is the constant term of the model,
- β_1 is the coefficient of the explanatory variable,
- ν is the parameter of the weighting function.

The estimation in this case is to determine the values of β_0, β_1, and ν that minimize the sum of squared residuals. This model is suitable for studying the effect of only one main factor on the change of the dependent variable.

มหาวิทยาลัยเทคโนโลยีสุรนารี

Case 2: Multivariate MIDAS

In the case where there is more than one explanatory variable (Andreou et al., 2010; Foroni et al., 2015), the model can be extended as follows:

$$Y_t = \beta_0 + \sum_{j=1}^{p} \beta_j \sum_{k=1}^{K} w_{j,k}(\nu_j) X_{j,t-k/m} + \varepsilon_t,$$

where $j = 1, 2, \ldots, p$ represents the index of each explanatory variable, in which each variable can have different numbers of lag terms, data frequencies, and weighting function structures.

The parameter estimation in this case requires the estimation of β_0, β_j, and ν_j for all variables simultaneously under the principle of minimizing the sum of squared residuals. This type of model is suitable for cases where the dependent variable is influenced by multiple factors simultaneously, which helps improve the ability to explain the data more effectively.

Case 3: MIDAS–AR

In some cases, the dependent variable may exhibit persistence from its own past values, meaning that the current value is also influenced by lagged values of the dependent variable. Therefore, the MIDAS model is extended by including autoregressive terms (Clements and Galvão, 2008; Ghysels et al., 2007), resulting in the MIDAS–AR model as follows:

$$Y_t = \beta_0 + \sum_{r=1}^{P} \phi_r Y_{t-r} + \sum_{j=1}^{p} \beta_j \sum_{k=1}^{K} w_{j,k}(\nu_j) X_{j,t-k/m} + \varepsilon_t,$$

where ϕ_r is the coefficient of the rth lagged dependent variable, and P is the order of the autoregressive process. The parameter estimation in this case requires the simultaneous estimation of β_0, ϕ_r, β_j, and ν_j, while still applying the Nonlinear Least Squares method in the same manner as the general MIDAS model.

Each type of MIDAS model has a different structure depending on the number of explanatory variables and temporal components. However, the parameter estimation process still relies on the same principle, namely finding the set of parameters that allows

the model to explain empirical data and provide forecasts with the highest efficiency.

In summary, the MIDAS family of models represents an important development in time series analysis and mixed-frequency forecasting because it allows high-frequency data to be directly incorporated with low-frequency dependent variables without the need for prior data aggregation. This helps reduce information loss and improve forecasting performance. These models have been continuously extended from the basic MIDAS framework to more complex forms, such as multivariate MIDAS and MIDAS–AR, which can effectively capture the influence of multiple factors together with the dynamics of past dependent variables. However, although MIDAS models have advantages in handling mixed-frequency data, they are still fundamentally based on the framework of linear relationships and assume that the model parameters remain constant throughout the entire period. This assumption may not be consistent with real-world data, which often exhibit structural changes, economic fluctuations, or market behaviors that vary over time. For this reason, researchers have developed approaches that can better accommodate changes in data conditions by integrating the MIDAS framework with models that allow parameters to vary according to different regimes or system states. This has led to the development of the Regime-Switching MIDAS model, which will be discussed in the next section.

2.6 Regime-Switching Model

2.6.1 Regime Concept in Financial Markets

In the context of financial markets, asset prices and returns often exhibit behavior that changes according to economic conditions, monetary policy, or geopolitical events. As a result, the statistical characteristics of time series data do not remain constant under a single structural regime, but instead display structural changes over different periods, referred to as “regimes” or states (Hamilton, 1989). In each regime, the parameters of the system, such as the mean (μ) or variance (σ^2), may differ significantly. Mathematically, a

state variable S_t can be defined to represent the regime at time t as follows:

$$S_t \in \{1, 2, \ldots, \mathcal{M}\},$$

where $\mathcal{M}$ is the number of specified regimes. For example, when $\mathcal{M} = 2$, the regimes may represent a normal state and a crisis state. Understanding this concept helps overcome the limitations of traditional linear models, which generally assume time-invariant parameters that remain constant throughout the entire period, an assumption that is often inconsistent with the reality of financial markets.

2.6.2 Threshold Regime-Switching Model

This is a model that determines the regime from the value of a directly observable variable. When the value of such variable is greater than or less than a threshold value, the system switches into another regime. This concept was systematically developed by Tong (1978) and Tong and Lim (1980).

An example of the basic form of the Threshold Autoregressive Model (TAR) can be written as follows:

$$Y_t = \begin{cases} \alpha_1 + \phi_1 Y_{t-1} + \epsilon_t, & q_t \leq c, \\ \alpha_2 + \phi_2 Y_{t-1} + \epsilon_t, & q_t > c, \end{cases}$$

where q_t is the threshold variable, and c is the threshold value.

For this study, the Threshold Regime-Switching Model can be divided into two forms as follows:

1. **Threshold MIDAS**

$$Y_t = \begin{cases} \beta_{0,1} + \sum_{j=1}^{p} \beta_{j,1} \sum_{k=1}^{K} w_{j,k}(\nu_{j,1}) X_{j,t-k/m} + \varepsilon_t, & Y_{t-1} \leq c, \\ \beta_{0,2} + \sum_{j=1}^{p} \beta_{j,2} \sum_{k=1}^{K} w_{j,k}(\nu_{j,2}) X_{j,t-k/m} + \varepsilon_t, & Y_{t-1} > c. \end{cases}$$

มหาวิทยาลัยเทคโนโลยีสุรนารี

2. **Threshold MIDAS–AR**

$$Y_t = \begin{cases} \beta_{0,1} + \sum_{r=1}^{P} \phi_{r,1} Y_{t-r} + \sum_{j=1}^{p} \beta_{j,1} \sum_{k=1}^{K} w_{j,k}(\nu_{j,1}) X_{j,t-k/m} + \varepsilon_t, & Y_{t-1} \leq c, \\ \beta_{0,2} + \sum_{r=1}^{P} \phi_{r,2} Y_{t-r} + \sum_{j=1}^{p} \beta_{j,2} \sum_{k=1}^{K} w_{j,k}(\nu_{j,2}) X_{j,t-k/m} + \varepsilon_t, & Y_{t-1} > c. \end{cases}$$

where

Y_{t-1} is the threshold variable,

c is the threshold value, and

d is the delay parameter.

The advantage of this type of model is that it is easy to interpret and suitable for data with clear turning points, such as when prices break through support or resistance levels. However, an important limitation is that an appropriate threshold value must be specified, and the regime switching occurs discontinuously.

2.6.3 Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) is a regime-switching model in which the regime or system state is defined as a hidden state that cannot be directly observed, but can be inferred from the behavior of the observed data. The fundamental concept of this model was developed by Baum and Petrie (1966), and was later widely applied in econometrics and finance, especially in the work of Hamilton (1989), which applied the Markov switching concept to explain business cycles.

Unlike the Threshold Regime-Switching Model, which determines the regime from directly observable variables, the HMM assumes that regimes evolve according to a Markov process. That is, the current state depends only on the state from the previous period, which can be written as follows:

$$P(S_t = j \mid S_{t-1} = i) = p_{ij},$$

where

- S_t is the state of the system at time t,
- i, j are the regime indices,
- p_{ij} is the probability that the system changes from state i to state j.

The transition probability matrix for the case of two regimes can be written as follows:

$$P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix},$$

where

- p_{11} is the probability that the system remains in regime 1,
- p_{12} is the probability that the system changes from regime 1 to regime 2,
- p_{21} is the probability that the system changes from regime 2 to regime 1,
- p_{22} is the probability that the system remains in regime 2.

The sum of probabilities in each row must be equal to 1.

For this study, two regimes are specified by using gold returns and gold price volatility as important information for estimating the hidden states of the gold market in order to classify normal market periods and high-volatility market periods before integrating the results with the MIDAS model. The Hidden Markov Model combined with the MIDAS framework can be divided into two forms as follows:

1. **HMM–MIDAS**

$$Y_t = \begin{cases} \beta_{0,1} + \sum_{j=1}^{p} \beta_{j,1} \sum_{k=1}^{K} w_{j,k}(\nu_{j,1}) X_{j,t-k/m} + \varepsilon_t, & S_t = 1, \\ \beta_{0,2} + \sum_{j=1}^{p} \beta_{j,2} \sum_{k=1}^{K} w_{j,k}(\nu_{j,2}) X_{j,t-k/m} + \varepsilon_t, & S_t = 2. \end{cases}$$

2. **HMM–MIDAS–AR**

$$Y_t = \begin{cases} \beta_{0,1} + \sum_{r=1}^{P} \phi_{r,1} Y_{t-r} + \sum_{j=1}^{p} \beta_{j,1} \sum_{k=1}^{K} w_{j,k}(\nu_{j,1}) X_{j,t-k/m} + \varepsilon_t, & S_t = 1, \\ \beta_{0,2} + \sum_{r=1}^{P} \phi_{r,2} Y_{t-r} + \sum_{j=1}^{p} \beta_{j,2} \sum_{k=1}^{K} w_{j,k}(\nu_{j,2}) X_{j,t-k/m} + \varepsilon_t, & S_t = 2, \end{cases}$$

where

- S_t is the hidden state at time t,
- p_{ij} is the transition probability from state i to state j,
- $\beta_{0,s}$ is the constant term under regime s,
- $\beta_{j,s}$ is the coefficient of the explanatory variable under regime s,
- $\phi_{r,s}$ is the autoregressive coefficient under regime s,
- $\nu_{j,s}$ is the weighting function parameter under regime s.

An important advantage of the HMM is its flexibility in adapting to market conditions without requiring a predefined threshold value. This allows the model to capture changes in unobserved states through a probabilistic process. However, the model is complex in terms of parameter estimation, and the results are sensitive to the specification of initial parameters, which may affect model stability.

Although the HMM is effective in identifying temporal regime switching, relying on only a small number of variables may not be sufficient to explain highly complex market conditions. Therefore, machine learning techniques such as the Self-Organizing Map (SOM) have recently been applied to cluster multivariate data and identify market patterns through dimensionality reduction before entering the HMM process. This approach helps improve forecasting accuracy and allows the determination of the number of regimes to be more consistent with the actual behavior of the data.

2.7 Self-Organizing Map

Self-Organizing Map (SOM) is an unsupervised learning technique developed by Teuvo Kohonen during the 1980s. Its objective is to cluster highly complex data and transform multidimensional data into a two-dimensional map that can be easily interpreted while preserving the structural relationships of the original data as much as possible (Kohonen, 1982, 1990).

The main concept of SOM is to feed multivariate data into a network of processing units (neurons) arranged on a two-dimensional grid. Each unit contains a weight vector representing the characteristics of the data. When a new observation enters the system, SOM searches for the unit that is most similar to the observation, called the Best Matching Unit (BMU). Then, the weights of the BMU and neighboring units are adjusted to move closer to the sample data. This process is repeatedly performed until systematic clustering of the data is achieved.

Let the input vector at time t be written as

$$x_t = (x_{1t}, x_{2t}, \ldots, x_{nt})^T,$$

where

- x_t is the input vector at time t,
- $x_{1t}, x_{2t}, \ldots, x_{nt}$ are the values of variables 1 to n at time t,
- n is the number of variables used as input data.

The BMU search uses the minimum Euclidean distance principle as follows:

$$c = \arg\min_i \|x_t - w_i\|,$$

where w_i denotes the weight vector associated with unit i. The weight vector of each unit is then updated according to

$$w_i(t+1) = w_i(t) + \eta(t) h_{ci}(t) \left[x_t - w_i(t)\right],$$

มหาวิทยาลัยเทคโนโลยีสุรนารี

where $\eta(t)$ is the learning rate and $h_{ci}(t)$ is the neighborhood function satisfying

$$0 \leq h_{ci}(t) \leq 1, \quad \text{with} \quad h_{cc}(t) = 1.$$

The learning rate is a parameter that determines the magnitude of the weight update of the SOM network in each learning iteration. In other words, it controls how much the processing units move toward the new input data. When $\eta(t)$ has a high value, the weight adjustment becomes larger, causing the network to learn faster. In contrast, when $\eta(t)$ has a low value, the weight adjustment becomes smaller, resulting in slower but more stable learning. In financial market research, SOM plays an important role in identifying market behavior patterns, such as low-volatility periods, high-volatility periods, or trend transition periods. These characteristics may not be clearly observable from raw data using only a single variable. In this study, SOM is used as a preprocessing step before estimating the Hidden Markov Model (HMM) in order to help cluster the data and extract structural patterns of the gold market from important variables, namely gold returns and gold price volatility. The SOM results are then used as supplementary information for estimating the hidden states of the HMM in order to improve the clarity of regime classification.

After regime classification, the obtained states are incorporated into the MIDAS–AR model for forecasting the dependent variable. The main model of this study is the SOM–HMM–MIDAS–AR Model, which is a hybrid model combining artificial intelligence techniques, hidden state models, and mixed-frequency data models.

The hybrid SOM–HMM–MIDAS–AR model can be written in the following general form:

$$Y_t = \begin{cases} \beta_{0,1} + \sum_{r=1}^{P} \phi_{r,1} Y_{t-r} + \sum_{j=1}^{p} \beta_{j,1} \sum_{k=1}^{K} w_{j,k}(\nu_{j,1}) X_{j,t-k/m} + \varepsilon_t, & S_t = 1, \\ \beta_{0,2} + \sum_{r=1}^{P} \phi_{r,2} Y_{t-r} + \sum_{j=1}^{p} \beta_{j,2} \sum_{k=1}^{K} w_{j,k}(\nu_{j,2}) X_{j,t-k/m} + \varepsilon_t, & S_t = 2, \end{cases}$$

where S_t is the hidden state of the system at time t, which is estimated from the Hidden Markov Model using data clustered by the Self-Organizing Map (SOM).

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.8 Information Criteria

In statistical and time series modeling, model selection is an important step because a model with too many parameters may lead to overfitting, while a model that is too simple may not be able to adequately explain the structure of the data (underfitting). Therefore, information criteria have been developed to compare multiple models under the same dataset by considering the balance between the ability to explain the data and the complexity of the model (Burnham and Anderson, 2002; Box et al., 2015).

The most commonly used information criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which can be applied to various types of models, such as linear models, ARIMA time series models, regression models, and other probabilistic models (Hamilton, 1994; Brockwell and Davis, 2016).

2.8.1 Akaike Information Criterion (AIC)

The AIC was developed by Hirotugu Akaike (1974) based on Information Theory in order to estimate the information loss when a model is used to represent the true process (Akaike, 1974).

$$AIC = -2\ln(L) + 2k,$$

where

- L is the maximum likelihood of the model,
- k is the number of parameters to be estimated.

The interpretation of the AIC value is that a lower AIC indicates a more appropriate model when compared with other models fitted to the same dataset.

2.8.2 Bayesian Information Criterion (BIC)

The BIC, or Schwarz Information Criterion, was proposed by Gideon Schwarz in 1978 and developed based on the Bayesian approach (Schwarz, 1978).

$$BIC = -2\ln(L) + k\ln(n),$$

where

- L is the maximum likelihood of the model,
- k is the number of parameters to be estimated,
- n is the number of observations.

The interpretation of the BIC value is similar to that of the AIC. A model with a lower BIC value indicates a more appropriate model when compared under the same dataset. In practice, the BIC is suitable for research focusing on identifying an appropriate model structure and avoiding overfitting, including studies that require clear parameter interpretation. However, one limitation is that, in some cases, it may select a model that is too simple, which can reduce forecasting performance.

Both AIC and BIC have the same objective, namely model comparison and model selection, but the two criteria emphasize different aspects. The AIC is more suitable for studies focusing on forecasting accuracy, whereas the BIC is more suitable for studies emphasizing model parsimony and theoretical interpretation (Lütkepohl, 2005). Therefore, in applied mathematics and time series research, both AIC and BIC are commonly considered together. If a model provides low values for both criteria in the same direction, it increases confidence that the model is appropriate in terms of both performance and complexity. Nevertheless, they should be used together with residual diagnostics and actual forecasting performance in order to achieve a more complete model selection process (Box et al., 2015).

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.9 Forecasting Performance Evaluation

Evaluating the performance of forecasting models is essential for assessing the accuracy and reliability of predictions. In this study, three widely used statistical measures are employed: Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Error (MAE). These indicators provide complementary perspectives on model performance.

2.9.1 Mean Absolute Error

MAE calculates the average absolute deviation between the predicted values $\left(\hat{Y}_t\right)$ and the actual values (Y_t), treating all errors equally. It is straightforward to interpret and provides practical insights into forecasting accuracy. The formula is defined as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \hat{Y}\right|,$$

where Y_t represents the actual value and $\hat{Y}_t$ the forecast value at time t.

2.9.2 Mean Absolute Percentage Error

MAPE expresses forecast accuracy as a percentage error by comparing the predicted values $\left(\hat{Y}_t\right)$ with the actual values (Y_t). This measure is particularly useful for comparing forecasting performance across datasets with different scales. The formula is defined as:

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right|.$$

2.9.3 Root Mean Square Error

RMSE measures the deviation between the predicted values $\left(\hat{Y}_t\right)$ and the actual values (Y_t), placing greater weight on large errors. Consequently, RMSE is highly sensitive to severe deviations. The formula is defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \hat{Y}_t\right)^2}.$$

By employing all three measures together, the evaluation offers a comprehensive assessment of model performance—capturing overall error magnitude (RMSE), relative percentage error (MAPE), and uniformly weighted average error (MAE).

2.10 Diebold–Mariano Test: DM-Test

The comparison of forecasting model performance should not rely only on average error measures such as RMSE, MAE, or MAPE because these values represent only the overall magnitude of forecasting errors and cannot clearly indicate whether the differences between models are statistically significant. Therefore, forecasting research commonly applies the Diebold–Mariano test (DM-Test), developed by Francis X. Diebold and Roberto S. Mariano in 1995, to examine whether two forecasting models have significantly different predictive accuracy (Diebold and Mariano, 1995).

The principle of this test is to compare the forecast loss of two models, such as Model 1 and Model 2. Let the actual value at time t be denoted by y_t, and the forecasts from each model be denoted by $\hat{y}_{1,t}$ and $\hat{y}_{2,t}$. The forecast errors can be calculated as follows:

$$e_{1,t} = y_t - \hat{y}_{1,t} \qquad e_{2,t} = y_t - \hat{y}_{2,t}.$$

Next, define a loss function $L(\cdot)$, such as squared error loss or absolute error loss, and define the loss differential as

$$d_t = L(e_{1,t}) - L(e_{2,t}).$$

If the mean of d_t is equal to zero, this indicates that the two models do not differ in forecasting accuracy. However, if the mean significantly differs from zero, it can be concluded that one model outperforms the other (Diebold and Mariano, 1995; Harvey et al., 1997).

The DM-Test statistic can be written as

$$DM = \frac{\bar{d}}{\sqrt{\widehat{Var}(\bar{d})}},$$

where

- $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t$ is the mean of the loss differential,
- $\widehat{Var}(\bar{d})$ is the estimated variance of the mean loss differential,

- T is the number of periods used for forecast evaluation.

Under the null hypothesis, the test statistic approximately follows the standard normal distribution.

Hypotheses

$H_0 : E(d_t) = 0$, meaning that the two models have no difference in forecasting ability.

$H_1 : E(d_t) \neq 0$, meaning that the two models have different forecasting ability.

If the p-value is less than the specified significance level, such as 0.05, the null hypothesis is rejected, and it can be concluded that the forecasting accuracy of the two models differs significantly. In the case where the DM statistic is positive, this indicates that Model 2 produces a smaller forecast loss than Model 1. In contrast, if the DM statistic is negative, this indicates that Model 1 performs better. However, this interpretation depends on the ordering of d_t.

The DM-Test has the advantage of directly comparing the forecasting accuracy of models based on forecast errors. In addition, it supports various forms of loss functions, such as Mean Squared Error (MSE) and Mean Absolute Error (MAE), making it flexible for application according to the characteristics of the data. Moreover, it is a standard method that has been widely accepted in forecasting and econometric research (Diebold and Mariano, 1995).

2.11 Related Research

Hamilton (1989) developed the Markov-Switching Model by allowing the parameters of an autoregressive process to change according to a discrete-state Markov chain. Maximum Likelihood Estimation combined with a nonlinear filter was employed to estimate the probabilities of being in different regimes. Using post–World War II U.S. Real GNP data, the study revealed a clear alternation between expansion and recession regimes, with recessions associated with a permanent decline in the level of GNP of approximately 3%.

Teräsvirta (1994) advanced the construction and evaluation of Smooth Transition Autoregressive (STAR) models, particularly the Logistic STAR (LSTAR) and Exponential STAR

(ESTAR) forms. The approach incorporates linearity testing to detect nonlinearity, selection of the delay parameter, and determination of the appropriate model type. Estimation is carried out via nonlinear least squares, with residual diagnostics applied for validation. Using both simulated data and empirical series such as Canadian lynx populations and German industrial production, the study demonstrated that STAR models provide a more flexible and continuous representation of regime transitions compared to linear or Threshold Autoregressive (TAR) models.

Hamilton and Susmel (1994) extended the regime-switching concept to financial volatility analysis by developing the Markov-Switching ARCH (SWARCH) model. This framework integrates Engle's(1982) ARCH model, which captures conditional heteroskedasticity, with Hamilton's (1989) MarkovSwitching mechanism, which models regimechanges in economicactivity. TheSWARCH modelallows volatility to shift flexibly across periods, enabling a better representation of high- and low-volatility regimes. Using weekly U.S. interest rate data, the study found that SWARCH outperformed the traditional ARCH model by more accurately capturing the alternation between high- and low-volatility states. These findings highlight the importance of incorporating regime structures into volatility modeling for a deeper understanding of financial market behavior and empirical financial econometrics.

Atsalakis and Valavanis (2009) reviewed intelligent forecasting methods in financial markets and highlighted SOM as an effective preprocessing and clustering tool for identifying hidden market structures before prediction. Their survey concluded that SOM-based hybrid models often outperform purely linear forecasting techniques in volatile financial environments.

Guérin and Marcellino (2013) introduced the Markov-Switching MIDAS (MS-MIDAS) model by integrating regime changes into the Mixed Data Sampling framework. The proposed model allows MIDAS parameters to vary across economic states governed by a latent Markov chain. Using U.S. and U.K. quarterly GDP data together with higher-frequency financial indicators, the study found that MS-MIDAS significantly improved both macroeconomic forecasting accuracy and real-time recession probability detection compared with standard MIDAS models.

Tsui, Xu, and Zhang (2013) conducted a study forecasting Singapore's GDP growth

one quarter ahead using mixed-frequency data. The research compared MIDAS regression with direct regression and time-averaging regression. The results demonstrated that the MIDAS model incorporating highfrequency stock return data significantly outperformed the other two approaches, with weekly stock returns providing the most accurate forecasts.

Nguyen and Le (2014) proposed a two-stage stock forecasting architecture combining Self-Organizing Map (SOM) with fuzzy Support Vector Machines. SOM was first used to partition market states into homogeneous clusters, after which fuzzy-SVM models were trained within each cluster. Experimental results on stock market datasets demonstrated superior forecasting accuracy compared with standalone machine learning models.

Barsoum and Stankiewicz (2015) proposed the Markov-Switching U-MIDAS (MS-U-MIDAS) model, which combines unrestricted MIDAS lag structures with regime-switching dynamics. The model was applied to U.S. GDP growth forecasting using a large mixed-frequency dataset. Empirical results showed that the MS-U-MIDAS specification produced similar or superior nowcasting and forecasting performance relative to conventional MIDAS approaches.

Ghysels and Qian (2019) investigated estimation techniques for the MIDAS regression model. Instead of relying on the commonly used but computationally complex Nonlinear Least Squares (NLS), they proposed using Ordinary Least Squares (OLS) combined with Polynomial Parameter Profiling. This approach incorporated a Beta polynomial weighting scheme to handle the lag parameters, simplifying linear estimation while preserving the efficiency and desirable properties of U-MIDAS. Their method enhanced the flexibility of MIDAS estimation and broadened its applicability to datasets with differing frequencies.

Fang et al. (2020) developed the GARCH-MIDAS model, which integrates short-term and longterm volatility components. The study applied variable selection using an Adaptive-Lasso penalty to identify the most influential macroeconomic and financial variables for long-term stock market volatility. Three key predictors were selected: housing starts, default spread, and realized volatility. Out-of-sample evaluation indicated that incorporating variable selection significantly improved forecasting accuracy, demonstrating the potential of the GARCH-MIDAS framework in predicting long-term stock market

volatility.

Tan et al. (2021) examined volatility modeling of Bitcoin (BTC) returns using a Time-Varying Transition Probability Markov-Switching GARCH (TV-MSGARCH) model with exogenous variables such as daily trading volume and Google search intensity. Comparisons with standard models, including GARCH, GJRGARCH, Threshold GARCH, and MSGARCH, showed that TV-MSGARCH with skewed and fat-tailed distributions provided the best in-sample fit. Moreover, the specification incorporating daily trading volume and Student-t error distribution achieved the highest out-of-sample forecasting accuracy. The study also identified time-varying transition probabilities and evaluated shortfall forecasts under both long and short trading positions, where TV-MSGARCH consistently outperformed benchmark models.

Cavicchioli (2021) proposed a closed-form matrix solution for computing higher-order moments and kurtosis in univariate Markov-Switching GARCH (MS-GARCH) models, along with asymptotic theory for the estimators of these quantities. These results facilitate normality testing and were validated through Monte Carlo simulations. The analytical framework was further applied to financial market data to identify periods of heightened volatility, such as during financial crises and the COVID-19 pandemic.

Kole and van Dijk (2023) employed the Markov-Switching Vector Autoregressive (MS-VAR) model to analyze the effects of economic and financial shocks under different regimes, such as recessions and bear markets. The MS-VAR framework effectivelycaptured both gradualadjustmentsand discrete regime shifts. The authors developed a comprehensive methodological approach, including regimeconditional first- and second-moment estimation and the construction of impulse response functions for both moments. Application to stock and bond market data revealed that forecasts of mean, volatility, and autocorrelations varied significantly across regimes. Moreover, shocks were found to exhibit strong nonlinearity, with more severe and persistent effects during downturn regimes.

Chan, Poon, and Zhu (2024) developed a Time-Varying Parameter MIDAS model for nowcasting U.S. real GDP. The framework allows regression coefficients to evolve over time rather than remain fixed. The study reported that the time-varying MIDAS model outperformed static MIDAS and benchmark regime-switching MIDAS models during periods

of structural instability.

Wang et al. (2024) investigated insurance demand forecasting in China using the Mixed Data Sampling (MIDAS) framework. The study utilized monthly data including the Consumer Confidence Index (CCI), Economic Policy Uncertainty (EPU) index, and Producer Price Index (PPI), along with quarterly data on Total Insurance Depth (TID). The results indicated that the MIDAS model provides highly accurate short-term forecasts, particularly when appropriate weighting functions and forecasting windows are applied. Moreover, the findings revealed that insurance demand is most sensitive to consumer confidence, especially during the COVID-19 period. This study underscores the importance of MIDAS as a tool for improving forecasting accuracy and supporting policy-making in the Chinese insurance sector.

Ma et al. (2024) proposed a hybrid model that combines Reserve Restricted Mixed Data Sampling (RR-MIDAS) with deep learning architectures, specifically CNN-LSTM, for forecasting stock price volatility using mixed-frequency data. The research emphasized handling macroeconomic variables and technical indicators with different frequencies and employed the Optuna framework to optimize parameter selection. Experimental results demonstrated that the proposed hybrid model significantly reduced forecasting errors, including RMSE, MAE, MSLE, MAPE, SMAPE, and QLIKE, compared to traditional models. On average, accuracy improvements ranged from 12% to 30%. These findings highlight the effectiveness and robustness of integrating RR-MIDAS with CNN-LSTM in forecasting stock price volatility under mixed-frequency conditions.

Although prior studies have separately demonstrated the forecasting capability of MIDAS models for mixed-frequency data and SOM models for nonlinear clustering, empirical evidence directly integrating SOM-based regime classification with MIDAS forecasting remains limited. Existing regime MIDAS studies mainly rely on Markov-switching mechanisms, while SOM offers a data-driven and nonparametric alternative for identifying hidden regimes. Therefore, combining SOM with MIDAS may provide a flexible framework capable of capturing nonlinear market states and improving forecasting performance under structural changes.

CHAPTER III

RESEARCH METHODOLOGY

This chapter presents the research methodology used in this study, organized into two main components. The research framework is divided into the development of appropriate models and the evaluation of model performance.

3.1 Phase I: Model Development

The model development in this study begins with the ARIMA model, which is a widely recognized linear time series model with a clear theoretical foundation. Therefore, it is used as the benchmark model for comparing the performance of the models developed in this study. Subsequently, the models were further developed to accommodate important characteristics of financial time series data discussed in Chapter 2, including mixed-frequency data, nonlinear relationships, and regime switching over time.

Under this framework, a sequence of models was developed, namely MIDAS, MIDAS-AR, MIDAS-AR-Threshold, MIDAS-AR-HMM, and MIDAS-AR-HMM-SOM. Each model was designed to reflect different characteristics of the data at varying levels, ranging from linear models to highly flexible models. The overall process of model development is illustrated in Figure 3.1.

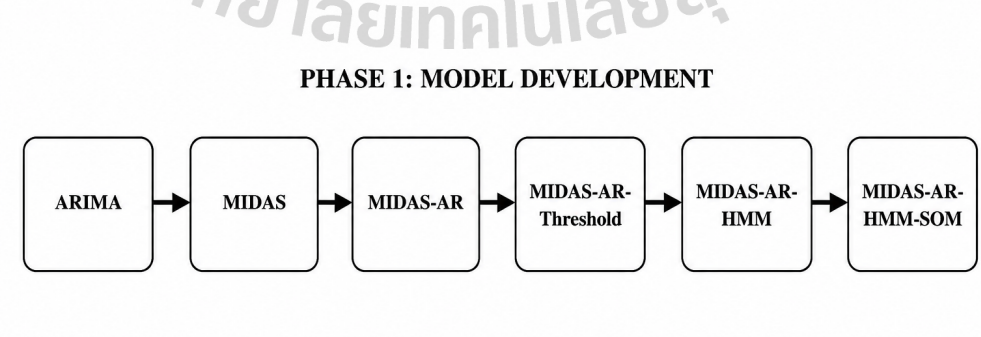


Figure 3.1 Phase I: Model Development.

3.2 Phase II: Model Performance Evaluation

The evaluation of model performance was conducted under the framework of out-of-sample forecasting in order to assess the applicability of the models to real-world data. The process in this phase consists of several important steps, including data preparation, data splitting, program development, and forecast performance evaluation.

3.2.1 Research Data

This study uses mixed-frequency time series data, in which Gold Price is used as the dependent variable in monthly form, while the independent variables, namely Oil Price and the U.S. Dollar Index, are in weekly form. The selection of these variables is based on financial economic theory. Oil prices reflect inflation conditions and production costs, while the U.S. Dollar Index has an inverse relationship with gold prices in the global market. Therefore, the use of high-frequency data together with low-frequency data is consistent with the concept of the MIDAS model, which can utilize additional information to improve forecasting performance. The data used in this study were collected from 2001 to 2025 in order to cover both normal economic periods and periods of high volatility, such as financial crises and periods of market uncertainty. This allows the models to learn the behavior of the data under different conditions.

The data sources are as follows:

- **Gold Price**

  The gold price data are represented in U.S. dollars per troy ounce.

  Yahoo Finance website: `https://finance.yahoo.com/quote/GC=F/history`

- **Oil Price**

  The oil price data are represented in U.S. dollars per barrel.

  Yahoo Finance website: `https://finance.yahoo.com/quote/CL=F/history`

- **U.S. Dollar Index**

  The U.S. Dollar Index is used as an indicator of the strength of the U.S. dollar in the global financial market.

Yahoo Finance website:

`https://finance.yahoo.com/quote/DX-Y.NYB/history`

All data were examined for completeness and consistency of the time periods before being used in the analysis.

3.2.2 Data Splitting

The time series data were divided into two parts, namely the training set and the testing set, for parameter estimation and the evaluation of model performance in out-of-sample forecasting. In this study, three data splitting ratios, namely 80:20, 75:25, and 70:30, were considered in order to compare forecasting results under different data conditions.

In the in-sample estimation stage, the training set was used to estimate model parameters and learn the structure of the time series data. The use of sufficiently long historical data allows the models to capture the behavior of the data under various economic conditions, including regime switching characteristics. In the out-of-sample forecasting stage, the testing set was used to evaluate the forecasting ability of the models on unseen data, reflecting the applicability of the models to real-world situations.

The forecasting results obtained from each data splitting ratio were compared to examine the consistency and stability of the models under different data conditions. If the forecasting results show similar patterns across different data splitting ratios, this indicates the reliability and practical applicability of the models. On the other hand, differences in forecasting performance were used to discuss the limitations and sensitivity of the models to data splitting conditions.

3.2.3 Program Development

For this study, the development and estimation of the models were conducted using the Python programming language, with Jupyter Notebook employed as the computational environment under the Windows 11 operating system and an Intel Core i5-13500H processor.

3.2.4 Forecast Performance Evaluation

The performance of the models was evaluated using forecasting error measures, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE), which reflect the magnitude of the differences between the actual values and the forecasted values.

In addition, information criteria, including the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), were used to evaluate the structural suitability of the models by considering both forecasting accuracy and model complexity.

From a statistical perspective, the comparison of forecasting accuracy among the models was further examined using the Diebold–Mariano (DM) test in order to determine whether the differences in forecasting errors between the models were statistically significant.

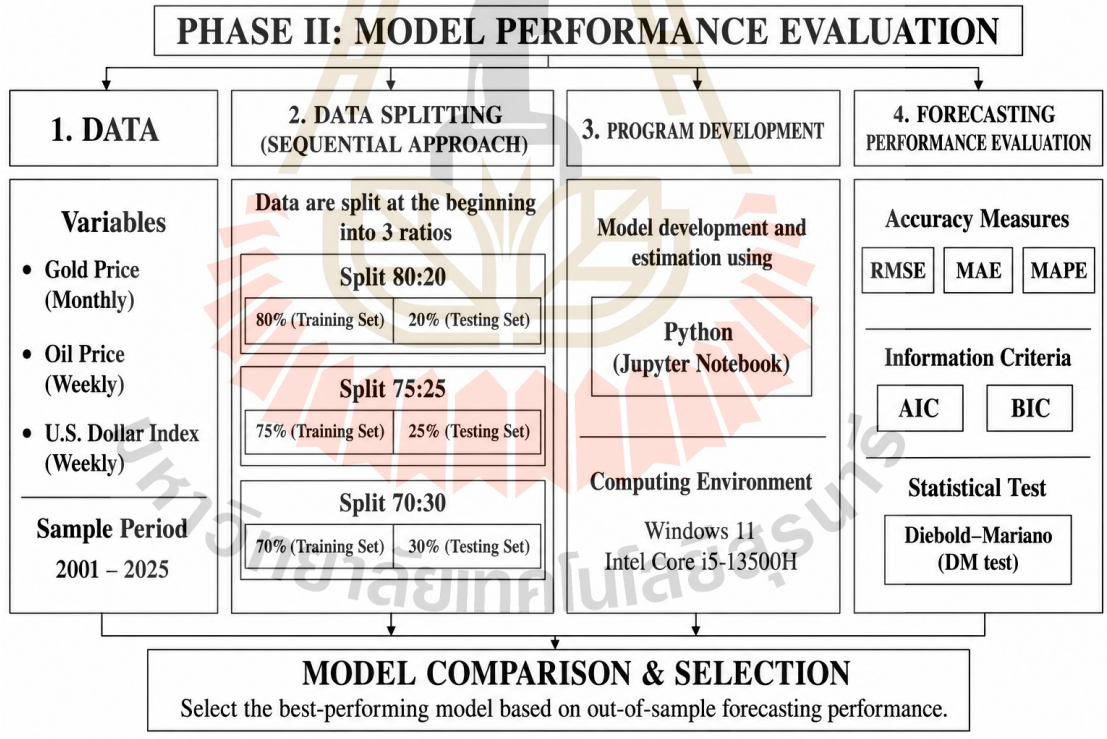


Figure 3.2 Phase II: Model Performance Evaluation.

CHAPTER IV

RESULTS

4.1 Overview of Research Results

This chapter presents the results of the development and performance evaluation of the models used for gold price forecasting. The models considered in this study include MIDAS, MIDAS-AR, MIDAS-AR-Threshold, MIDAS-AR-HMM, and MIDAS-AR-HMM-SOM, each evaluated using three lag lengths: $K = 4$, $K = 8$, and $K = 12$. In addition, the ARIMA model is employed as a baseline model for performance comparison.

The performance of the models is evaluated using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Ljung–Box test, and forecasting error measures, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Furthermore, the robustness of the models is examined under different Train:Test ratios, namely 70:30, 75:25, and 80:20.

4.2 Results of ARIMA Model Analysis

The ARIMA model was used as the baseline model for comparison with the other models in this study. Monthly gold price data were used to construct the model, and the data were divided into training and testing sets under different Train ratios, namely 70:30, 75:25, and 80:20. Prior to model construction, the stationarity of the data was examined using the Augmented Dickey–Fuller test (ADF test). The results indicated that the original series was non-stationary, as the p-value was greater than 0.05. Therefore, first-order differencing was applied before the data were used in the model-building process. Subsequently, model order selection in this study was performed using the Auto ARIMA function in Python, which selects the most appropriate model based on the minimum AIC criterion.

4.2.1 Stationarity Testing

Table 4.1 Results of the ADF stationarity test before and after first-order differencing.

Train:Test	Data	ADF Statistic	p-value	Conclusion
70:30	Original Series	-1.474	0.547	Non-Stationary
70:30	First Difference	-3.646	0.005	Stationary
75:25	Original Series	-1.268	0.644	Non-Stationary
75:25	First Difference	-3.708	0.004	Stationary
80:20	Original Series	-0.835	0.809	Non-Stationary
80:20	First Difference	-3.677	0.004	Stationary

Table 4.1 presents the results of the ADF test, which was used to examine the stationarity of the data under different Train:Test ratios. The results show that the original gold price series has p-values greater than 0.05 in all cases, indicating that the data are non-stationary. After applying first-order differencing, the p-values are less than 0.05 in all cases. Therefore, the first-order differenced data can be used for constructing the ARIMA model.

4.2.2 ARIMA Model Selection Results

Table 4.2 presents the results of ARIMA model selection under different Train ratios.

Table 4.2 ARIMA Model Selection Results.

Train:Test	Selected Model	AIC	BIC
70:30	ARIMA(0,1,1)	2267.155	2277.182
75:25	ARIMA(0,1,1)	2425.280	2435.515
80:20	ARIMA(0,1,1)	2601.736	2612.165

Based on the Auto ARIMA procedure implemented in Python, the results indicate that the ARIMA(0,1,1) model is the most appropriate model under all Train ratios.

4.2.3 Residual Diagnostic Checking

The residuals of the model were examined using the Ljung–Box test, which was employed to assess the white noise property of the residuals.

Table 4.3 Ljung–Box Test Results for the ARIMA Model.

Train:Test	Model	Ljung–Box p-value	Conclusion
70:30	ARIMA(0,1,1)	0.295	Pass
75:25	ARIMA(0,1,1)	0.307	Pass
80:20	ARIMA(0,1,1)	0.402	Pass

The results indicate that the residuals of the ARIMA model have p-values greater than 0.05 in all cases, suggesting that the residuals do not exhibit statistically significant autocorrelation.

4.2.4 Performance Evaluation Results of the ARIMA Model on the Test Set

To evaluate the performance of the model, forecasting error measures, including RMSE, MAE, and MAPE, were calculated using the test dataset.

Table 4.4 Forecasting Performance of the ARIMA Model.

Train:Test	Model	RMSE	MAE	MAPE
70:30	ARIMA(0,1,1)	787.544	553.646	22.733
75:25	ARIMA(0,1,1)	739.552	491.300	18.592
80:20	ARIMA(0,1,1)	607.538	387.759	14.082

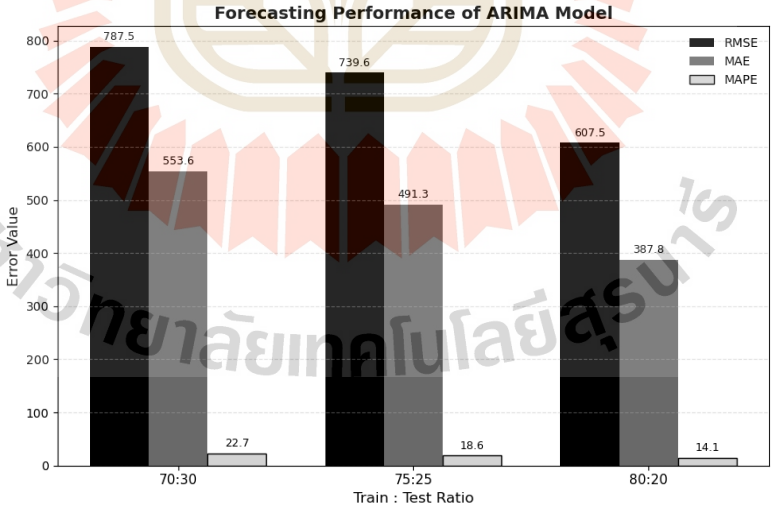


Figure 4.1 Forecasting Performance of ARIMA Model under Different Train:Test Ratios.

The results indicate that as the proportion of training data decreases, the forecasting errors of the ARIMA model tend to increase. Among the ARIMA models, the Train ratio of 80:20 provides the best forecasting performance.

Table 4.5 Parameter Estimates of the ARIMA(0,1,1) Model for the Train:Test Ratio of 80:20.

Parameter	Coefficient	Std. Error	p-value
Intercept	6.755	3.052	0.027
MA(1)	-0.152	0.042	0.000
Variance (σ^2)	3050.172	197.161	0.000

Table 4.5 presents the parameter estimates of the ARIMA(0,1,1) model for the Train ratio of 80:20. The results show that the MA(1) parameter is statistically significant at the 0.05 significance level, with a p-value less than 0.05. In addition, the Ljung–Box test yields a p-value of 0.402, indicating that the residuals of the model do not exhibit significant autocorrelation. The estimated ARIMA(0,1,1) model can be expressed as follows:

$$\Delta Y_t = 6.755 - 0.152\varepsilon_{t-1} + \varepsilon_t.$$

4.3 Results of the MIDAS Model

This section presents the forecasting results of the MIDAS model, which uses gold prices as the dependent variable and the DXY index together with oil prices as independent variables. The main advantage of the MIDAS model is its ability to incorporate variables with different frequencies into the estimation and forecasting process. In this study, the dependent variable, gold price, is observed at a monthly frequency, while the independent variables, the DXY index and oil prices, are observed at a weekly frequency. Three lag structures are considered: $K = 4$, $K = 8$, and $K = 12$.

4.3.1 Analysis of AIC and BIC Values for the MIDAS Model

Table 4.6 AIC and BIC Values of the MIDAS Model.

Train:Test	$K = 4$		$K = 8$		$K = 12$	
	AIC	BIC	AIC	BIC	AIC	BIC
70:30	1294.700	1312.200	1294.200	1311.700	1293.110	1310.610
75:25	1087.799	1104.022	1098.268	1114.491	1086.048	1102.270
80:20	876.970	891.630	884.475	899.135	880.730	895.397

The results indicate that the MIDAS model with the lowest AIC value differs across the Train:Test ratios. For the Train:Test ratios of 70:30 and 75:25, the model with $K = 12$

yields the lowest AIC value, whereas for the Train:Test ratio of 80:20, the model with $K = 4$ yields the lowest AIC value.

When considering the BIC values, the model with $K = 12$ also produces the lowest BIC value for the Train:Test ratios of 70:30 and 75:25. In contrast, for the Train:Test ratio of 80:20, the model with $K = 4$ yields the lowest BIC value.

4.3.2 Residual Diagnostic Checking

The residuals of the MIDAS model were examined using the Ljung–Box test to assess the white noise property of the residuals.

Table 4.7 Ljung–Box Test Results for the MIDAS Model.

Train:Test	$K = 4$	$K = 8$	$K = 12$
	Ljung–Box p-value	Ljung–Box p-value	Ljung–Box p-value
70:30	4.42×10^{-72}	3.98×10^{-72}	3.86×10^{-71}
75:25	2.34×10^{-58}	5.41×10^{-61}	2.55×10^{-55}
80:20	2.74×10^{-51}	2.41×10^{-52}	3.30×10^{-51}

The Ljung–Box test results indicate that the residuals of the MIDAS model exhibit autocorrelation in all cases.

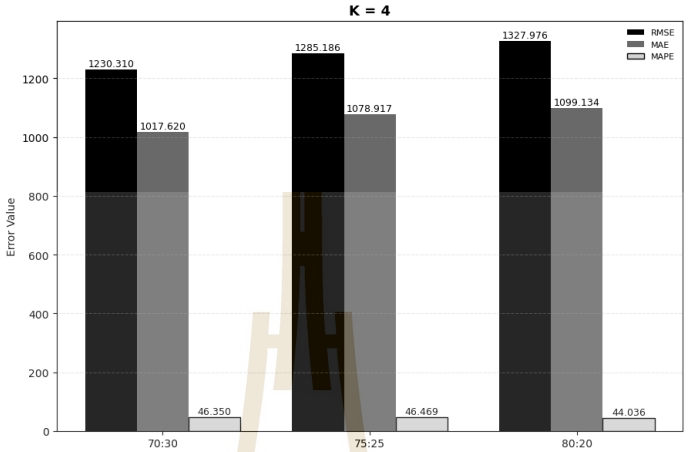
4.3.3 Performance Evaluation of the Model on the Test Dataset

Table 4.8 Performance Evaluation Results of the MIDAS Model under Different Train:Test Ratios.

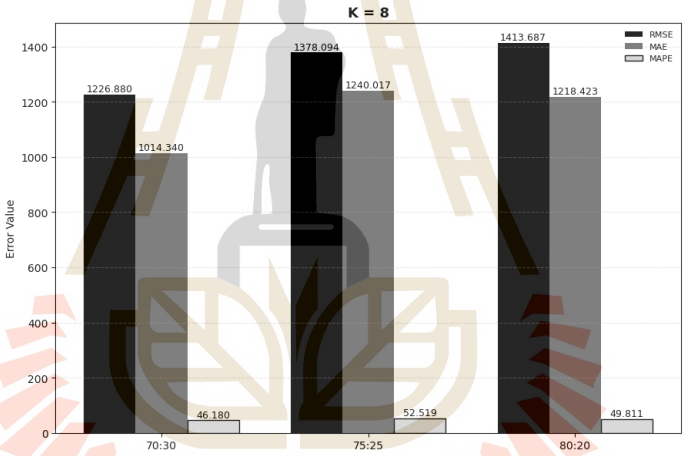
Train:Test	$K = 4$			$K = 8$			$K = 12$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
70:30	1230.310	1017.620	46.350	1226.880	1014.340	46.180	1219.480	1009.940	45.830
75:25	1285.186	1078.917	46.469	1378.094	1240.017	52.519	1270.270	1070.580	46.205
80:20	1327.976	1099.134	44.036	1413.687	1218.423	49.811	1370.329	1167.458	47.477

The results indicate that the MIDAS model yields relatively high forecasting errors in all cases. When comparing different lag lengths, the model with $K = 4$ provides the best forecasting performance for the Train:Test ratio of 80:20, with an RMSE of 1327.976, an MAE of 1099.134, and a MAPE of 44.036%.

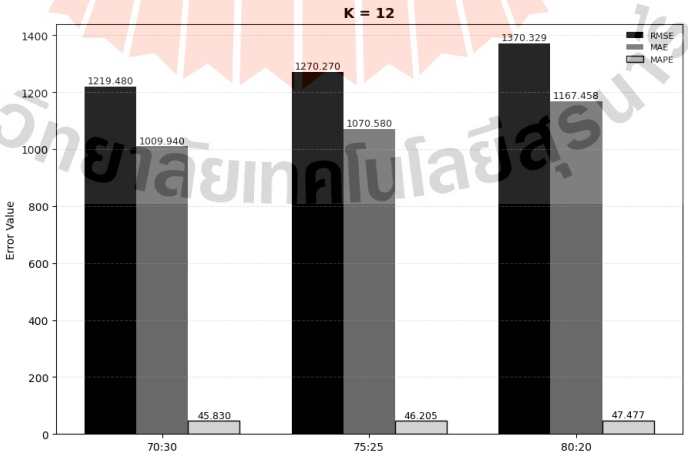
For the Train:Test ratios of 70:30 and 75:25, the model with $K = 12$ yields the lowest forecasting errors when compared with the cases of $K = 4$ and $K = 8$.



(a) $K = 4$



(b) $K = 8$



(c) $K = 12$

Figure 4.2 Forecasting Performance of the MIDAS Model under Different K Values and Train:Test Ratios.

4.3.4 The Selected MIDAS Model

Based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and forecasting error measures, including RMSE, MAE, and MAPE, the MIDAS model with $K = 4$ under the Train:Test ratio of 80:20 provides the best forecasting performance on the test dataset. In particular, this model yields the lowest MAPE value compared with the other cases. In addition, this case utilizes a larger proportion of training data. Therefore, the MIDAS model with $K = 4$ under the Train:Test ratio of 80:20 is selected as the representative model for presenting the structure of the MIDAS model.

The general form of the MIDAS model can be expressed as follows:

$$Y_t = \beta_0 + \beta_1 \sum_{k=1}^{4} w_{1,k}(\nu_1) X_{t-k/m}^{(DXY)} + \beta_2 \sum_{k=1}^{4} w_{2,k}(\nu_2) X_{t-k/m}^{(OIL)} + \varepsilon_t,$$

where Y_t denotes the gold price at time t, while $X^{(DXY)}$ and $X^{(OIL)}$ represent the DXY index and oil price variables, respectively. The terms $w_{1,k}(\nu_1)$ and $w_{2,k}(\nu_2)$ denote the lag weights in the MIDAS model.

In this study, the lag weights are estimated using the Exponential Almon Lag Polynomial. The estimated parameters of the MIDAS model that provides the best forecasting performance are presented in Appendix A.1.

Therefore, in the next section, the MIDAS-AR model is developed by incorporating an autoregressive component into the model structure for forecasting gold prices.

4.4 Results of the MIDAS-AR Model

This section presents the forecasting results of the MIDAS-AR model, which is an extension of the MIDAS model obtained by incorporating an autoregressive component into the MIDAS model structure.

4.4.1 Analysis of AIC and BIC Values for the MIDAS-AR Model

Table 4.9 shows that for the Train:Test ratios of 70:30 and 75:25, the MIDAS-AR model with $K = 4$ yields the lowest AIC and BIC values. In contrast, for the Train:Test

Table 4.9 AIC and BIC Values of the MIDAS-AR Model.

Train:Test	$K = 4$		$K = 8$		$K = 12$	
	AIC	BIC	AIC	BIC	AIC	BIC
70:30	1688.220	1714.960	1877.386	1904.124	1874.809	1901.560
75:25	1803.926	1831.219	1928.653	1955.946	2010.140	2037.397
80:20	1936.791	1964.603	2166.121	2193.933	1930.380	1958.158

ratio of 80:20, the model with $K = 12$ produces the lowest AIC and BIC values.

4.4.2 Residual Diagnostic Checking

Table 4.10 Ljung–Box Test Results for the MIDAS-AR Model.

Train:Test	$K = 4$	$K = 8$	$K = 12$
	Ljung–Box p-value	Ljung–Box p-value	Ljung–Box p-value
70:30	0.019	0.021	0.004
75:25	0.025	8.77×10^{-7}	1.8×10^{-4}
80:20	0.036	9.76×10^{-8}	0.055

Table 4.10 shows that although the MIDAS-AR model with $K = 4$ under the Train:Test ratios of 70:30 and 75:25 yields the lowest AIC and BIC values, the Ljung–Box test indicates that the residuals still exhibit autocorrelation, since the p-values are less than 0.05. In contrast, for the Train:Test ratio of 80:20 with $K = 12$, the p-value of the Ljung–Box statistic is greater than 0.05, indicating that the residuals do not exhibit statistically significant autocorrelation.

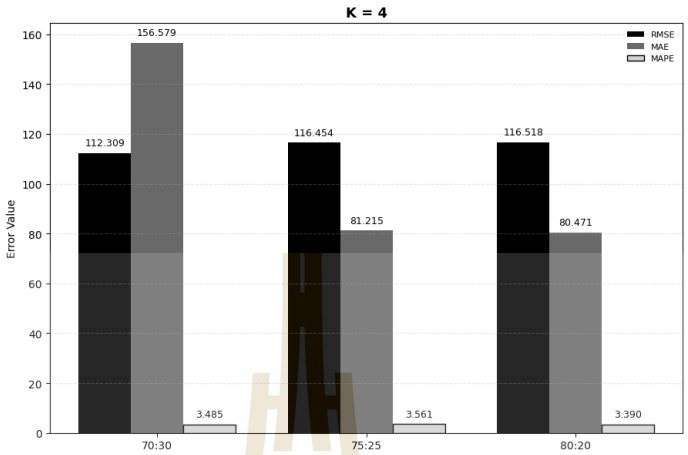
4.4.3 Performance Evaluation of the Model on the Test Dataset

Table 4.11 Performance Evaluation Results of the MIDAS-AR Model under Different Train:Test Ratios.

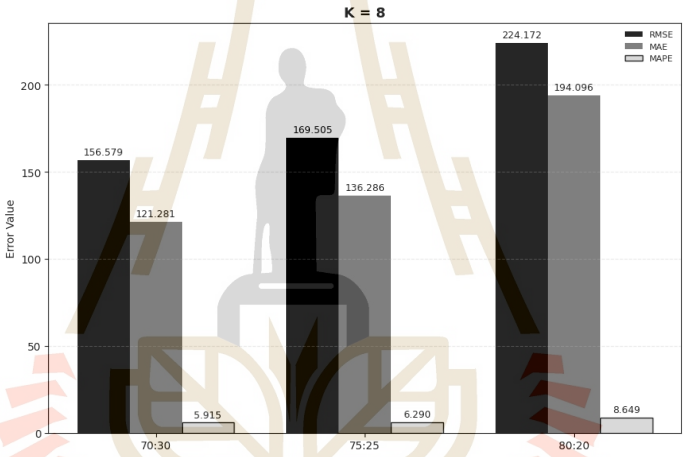
Train:Test	$K = 4$			$K = 8$			$K = 12$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
70:30	112.309	156.579	3.485	156.579	121.281	5.915	181.826	143.373	6.769
75:25	116.454	81.215	3.561	169.505	136.286	6.290	204.205	168.575	7.647
80:20	116.518	80.471	3.390	224.172	194.096	8.649	114.239	79.739	3.383

The performance evaluation results indicate that the MIDAS-AR model substantially reduces forecasting errors compared with the MIDAS model. In particular, the case

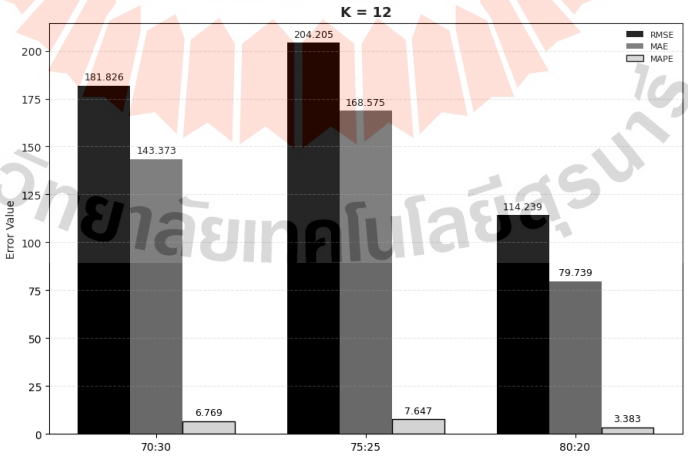
with a Train:Test ratio of 80:20 and $K = 12$ yields lower RMSE, MAE, and MAPE values.



(a) $K = 4$



(b) $K = 8$



(c) $K = 12$

Figure 4.3 Forecasting Performance of the MIDAS-AR Model under Different K Values and Train:Test Ratios.

4.4.4 The Selected MIDAS-AR Model

Based on the AIC, BIC, and forecasting error measures, including RMSE, MAE, and MAPE, the MIDAS-AR model with $K = 12$ under the Train:Test ratio of 80:20 provides the most appropriate forecasting performance in this study. Therefore, the MIDAS-AR model with $K = 12$ under this Train:Test ratio is selected as the representative model for presenting the structure of the MIDAS-AR model.

The general form of the MIDAS-AR model can be expressed as follows:

$$Y_t = \beta_0 + \phi_1 Y_{t-1} + \beta_1 \sum_{k=1}^{12} w_{1,k}(\nu_1) X_{t-k/m}^{(DXY)} + \beta_2 \sum_{k=1}^{12} w_{2,k}(\nu_2) X_{t-k/m}^{(OIL)} + \varepsilon_t,$$

where Y_t denotes the gold price at time t, and Y_{t-1} denotes the lagged gold price. The variables $X^{(DXY)}$ and $X^{(OIL)}$ represent the DXY index and oil price, respectively. The terms $w_{1,k}(\nu_1)$ and $w_{2,k}(\nu_2)$ denote the lag weights in the MIDAS-AR model, and ε_t represents the model error term. In this study, the lag weights are estimated using the Exponential Almon Lag Polynomial. The estimated parameters of the MIDAS-AR model that provides the best forecasting performance are presented in Appendix A.2.

In the next section, the MIDAS-AR-Threshold model is developed to investigate the performance of the model under regime changes.

4.5 Results of the MIDAS-AR-Threshold Model

This section presents the forecasting results of the MIDAS-AR-Threshold model, which incorporates regime-switching behavior by using a gold price threshold to divide the data into different regimes.

4.5.1 Analysis of AIC and BIC Values for the MIDAS-AR-Threshold Model

The information criteria results for the MIDAS-AR-Threshold model are presented in 4.12.

Table 4.12 AIC and BIC Values of the MIDAS-AR-Threshold Model.

Train:Test	$K = 4$		$K = 8$		$K = 12$	
	AIC	BIC	AIC	BIC	AIC	BIC
70:30	2293.925	2350.745	2293.184	2350.003	2286.862	2343.600
75:25	2452.783	2510.781	2453.081	2511.079	2443.262	2501.184
80:20	2628.081	2687.181	2631.032	2690.132	2620.818	2679.847

Table 4.12 shows that for the Train:Test ratios of 70:30, 75:25, and 80:20, the MIDAS-AR-Threshold model with $K = 12$ yields the lowest AIC and BIC values.

4.5.2 Residual Diagnostic Checking

Table 4.13 Ljung–Box Test Results for the MIDAS-AR-Threshold Model.

Train:Test	$K = 4$	$K = 8$	$K = 12$
	Ljung–Box p-value	Ljung–Box p-value	Ljung–Box p-value
70:30	0.089	0.072	0.039
75:25	0.067	0.051	0.039
80:20	0.029	0.038	0.056

Table 4.13 shows that although the MIDAS-AR-Threshold model with $K = 12$ under the Train:Test ratios of 70:30 and 75:25 yields the lowest AIC and BIC values, the Ljung–Box test indicates that the residuals still exhibit autocorrelation, since the p-value is 0.039, which is less than 0.05.

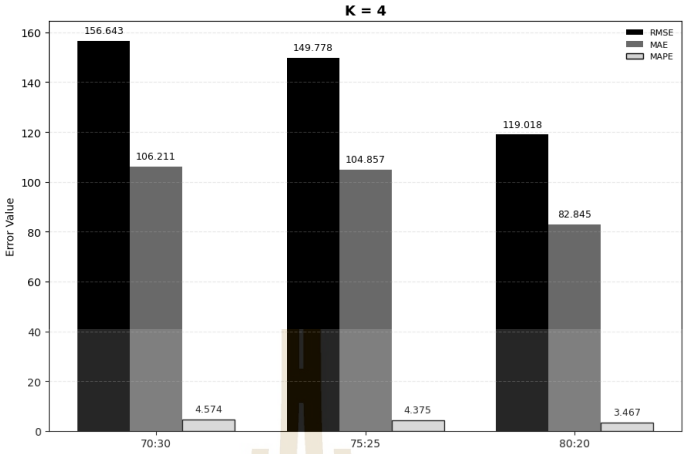
In contrast, for the Train:Test ratio of 80:20 with $K = 12$, the Ljung–Box statistic yields a p-value of 0.056, which is greater than 0.05, indicating that the residuals do not exhibit autocorrelation.

4.5.3 Performance Evaluation of the Model on the Test Dataset

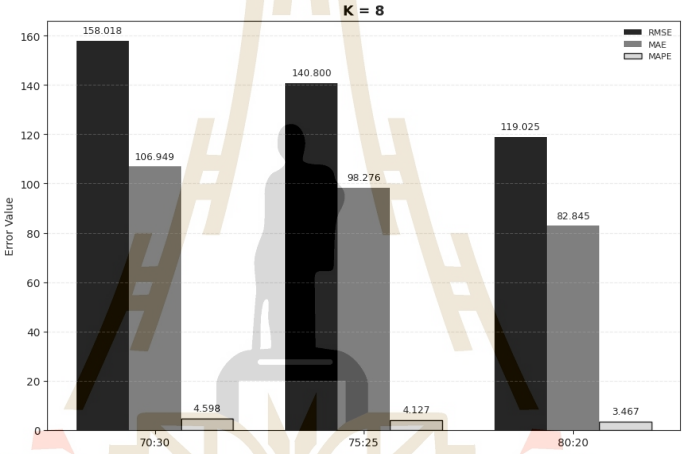
Table 4.14 Performance Evaluation Results of the MIDAS-AR-Threshold Model under Different Train:Test Ratios.

Train:Test	$K = 4$			$K = 8$			$K = 12$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
70:30	156.643	106.211	4.574	158.018	106.949	4.598	135.126	90.721	3.964
75:25	149.778	104.857	4.375	140.800	98.276	4.127	136.735	95.454	4.029
80:20	119.018	82.845	3.467	119.025	82.845	3.467	119.924	83.354	3.478

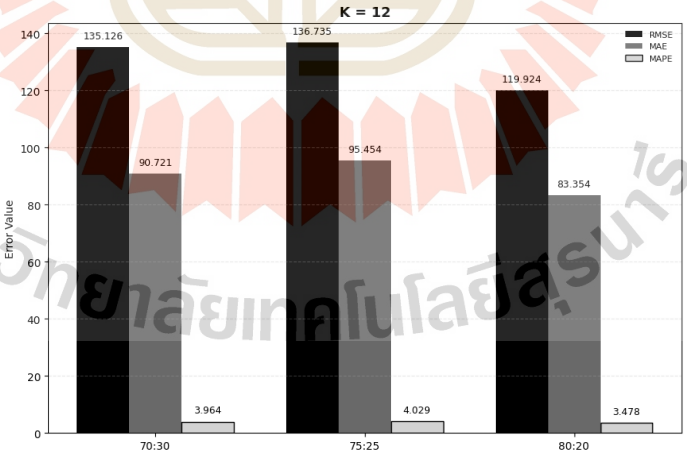
มหาวิทยาลัยเทคโนโลยีสุรนารี



(a) $K = 4$



(b) $K = 8$



(c) $K = 12$

Figure 4.4 Forecasting Performance of the MIDAS-AR-Threshold Model under Different K Values and Train:Test Ratios.

4.5.4 Regime Determination and the Selected MIDAS-AR-Threshold Model

Based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Ljung–Box test results, and forecasting error measures, including RMSE, MAE, and MAPE, the MIDAS-AR-Threshold model with $K = 12$ under the Train:Test ratio of 80:20 provides the most appropriate forecasting performance. This model yields the lowest AIC and BIC values, does not exhibit autocorrelation in the residuals, and produces relatively low forecasting errors compared with the other cases.

In this study, the MIDAS-AR-Threshold model uses the one-period lagged gold price (Y_{t-1}) as the threshold variable for regime classification, defined as

$$\text{Regime}_t = \begin{cases} 1, & Y_{t-1} \leq c, \\ 2, & Y_{t-1} > c, \end{cases}$$

where c denotes the threshold parameter estimated from the model estimation process.

Table 4.15 Threshold Value of the MIDAS-AR-Threshold Model.

Train:Test	K	Threshold	Regime 1	Regime 2
80:20	12	1171.3	$Y_{t-1} \leq c$	$Y_{t-1} > c$

The threshold value reported in Table 4.15 was selected as the value that minimizes the residual sum of squares (RSS) during the threshold estimation procedure.

The MIDAS-AR-Threshold model can then be expressed in the following general form:

$$Y_t = \begin{cases} \beta_{0,1} + \phi_1 Y_{t-1} + \\ \beta_{1,1} \sum_{k=1}^{12} w_{1,k}(\nu_{1,1}) X_{t-k/m}^{DXY} + \beta_{2,1} \sum_{k=1}^{12} w_{2,k}(\nu_{2,1}) X_{t-k/m}^{OIL} + \varepsilon_t, & Y_{t-1} \leq c, \\ \beta_{0,2} + \phi_2 Y_{t-1} + \\ \beta_{1,2} \sum_{k=1}^{12} w_{1,k}(\nu_{1,2}) X_{t-k/m}^{DXY} + \beta_{2,2} \sum_{k=1}^{12} w_{2,k}(\nu_{2,2}) X_{t-k/m}^{OIL} + \varepsilon_t, & Y_{t-1} > c, \end{cases}$$

where Y_t denotes the gold price at time t, and Y_{t-1} denotes the lagged gold price. The variables $X^{(DXY)}$ and $X^{(OIL)}$ represent the DXY index and oil price, respectively. The

terms $w_{1,k}$ and $w_{2,k}$ denote the lag weights in the MIDAS-AR-Threshold model $(i = 1, 2)$, and ε_t represents the model error term. The estimated parameters of the MIDAS-AR-Threshold model that provides the best forecasting performance are presented in Appendix A.3.

4.6 Results of the MIDAS-AR-HMM Model

This section presents the forecasting results of the MIDAS-AR-HMM model, which is an extension of the MIDAS-AR model by incorporating a Hidden Markov Model (HMM) to classify market regimes before estimating the MIDAS-AR model within each regime. In this study, regime classification is based on gold returns and the volatility of gold returns as observed variables, in order to capture the changing characteristics of the gold market over time.

4.6.1 Regime Classification Using the Hidden Markov Model

In this study, market regimes are not determined by a fixed threshold as in the MIDAS-AR-Threshold model. Instead, a Hidden Markov Model (HMM) is used to learn market states from the statistical characteristics of gold prices. The observed variables consist of gold returns and volatility, defined as

$$D_t = (r_t, \sigma_t),$$

where $r_t = \ln\left(\frac{Y_t}{Y_{t-1}}\right)$ denotes the logarithmic return of the gold price at time t, where Y_t represents the gold price at time t.

The variable σ_t denotes volatility, which is calculated using the rolling standard deviation of gold returns as follows:

$$\sigma_t = \sqrt{\frac{1}{m-1}\sum_{i=0}^{m-1}(r_{t-i} - \bar{r})^2},$$

where m denotes the rolling window, which is set to $m = 3$ in this study.

In this study, the Gaussian Hidden Markov Model is specified with two hidden states to represent different market regimes. The HMM learns the patterns in the data from the distribution of gold returns and volatility while considering the temporal continuity of regimes.

In addition, the HMM uses a Transition Probability Matrix to describe the probabilities of remaining in or switching between regimes over time, which can be expressed as follows:

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix},$$

where p_{ij} denotes the probability that the market transitions from regime i to regime j in the next period.

For this study, the estimated transition probability matrix, whose parameters are estimated from the training data, is given by

$$P = \begin{pmatrix} 0.977 & 0.023 \\ 0.257 & 0.743 \end{pmatrix}.$$

The estimated transition probability matrix shows that the probabilities on the main diagonal (diagonal entries) are relatively high, particularly $p_{11} = 0.977$, indicating that once the market is in Regime 1, it is highly likely to remain in the same regime in the following period. Similarly, $p_{22} = 0.743$ indicates that Regime 2 also exhibits persistence over time, although it has a higher probability of switching back to Regime 1 compared with Regime 1.

มหาวิทยาลัยเทคโนโลยีสุรนารี

4.6.2 Analysis of AIC and BIC Values for the MIDAS-AR-HMM Model

The information criteria results for the MIDAS-AR-HMM model are presented in Table 4.16.

Table 4.16 AIC and BIC Values of the MIDAS-AR-HMM Model.

Train:Test	$K = 4$		$K = 8$		$K = 12$	
	AIC	BIC	AIC	BIC	AIC	BIC
70:30	1668.592	1721.915	1697.447	1750.771	1665.408	1708.731
75:25	1783.724	1838.167	1813.046	1867.489	1771.150	1825.593
80:20	1907.645	1963.066	1934.989	1990.410	1894.908	1950.330

Table 4.16 shows that for the Train:Test ratios of 70:30, 75:25, and 80:20, the MIDAS-AR-HMM model with $K = 12$ yields the lowest AIC and BIC values.

4.6.3 Residual Diagnostic Checking

Table 4.17 Ljung–Box Test Results for the MIDAS-AR-HMM Model.

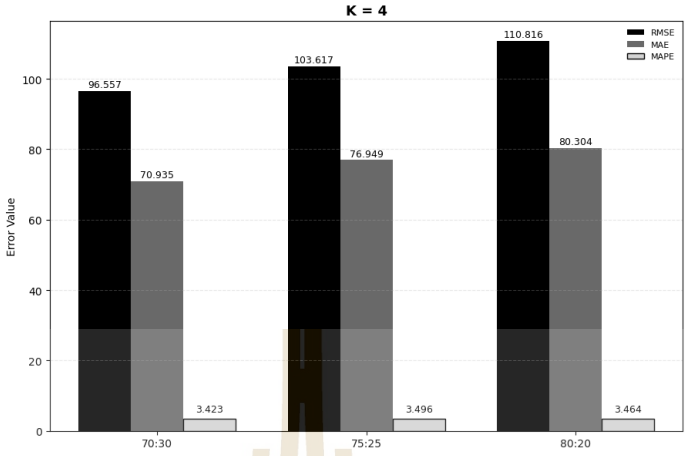
Train:Test	$K = 4$	$K = 8$	$K = 12$
	Ljung–Box p-value	Ljung–Box p-value	Ljung–Box p-value
70:30	0.960	0.954	0.950
75:25	0.041	0.056	0.099
80:20	0.714	0.651	0.645

Table 4.17 shows that for the Train:Test ratio of 75:25 with $K = 4$, the Ljung–Box test indicates that the residuals exhibit autocorrelation, since the p-value is less than 0.05.

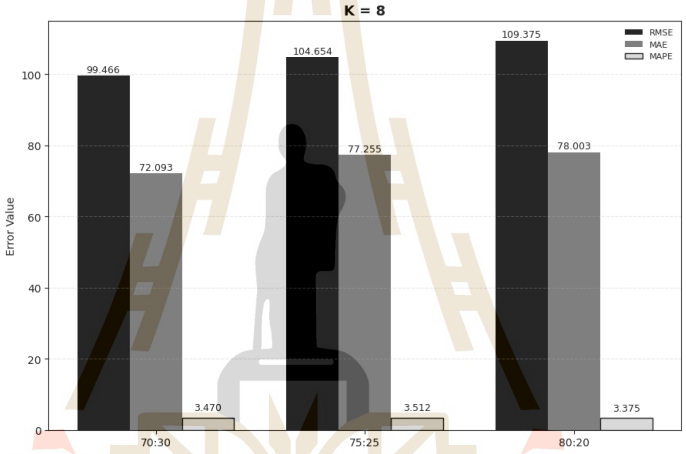
4.6.4 Performance Evaluation of the Model on the Test Dataset

Table 4.18 Performance Evaluation Results of the MIDAS-AR-HMM Model under Different Train:Test Ratios.

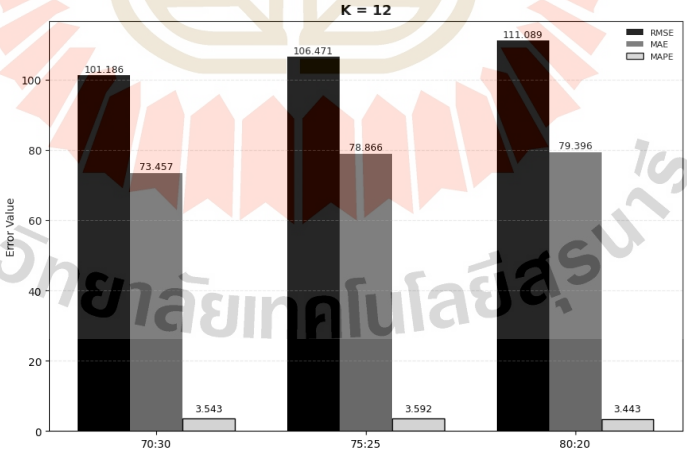
Train:Test	$K = 4$			$K = 8$			$K = 12$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
70:30	96.557	70.935	3.423	99.446	72.093	3.470	101.186	73.457	3.543
75:25	103.617	76.949	3.496	104.654	77.255	3.512	106.471	78.866	3.592
80:20	110.816	80.304	3.464	109.375	78.0039	3.375	111.089	79.396	3.443



(a) $K = 4$



(b) $K = 8$



(c) $K = 12$

Figure 4.5 Forecasting Performance of the MIDAS-AR-HMM Model under Different K Values and Train:Test Ratios.

4.6.5 The Selected MIDAS-AR-HMM Model

Based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and forecasting error measures, including RMSE, MAE, and MAPE, the MIDAS-AR-HMM model provides good forecasting performance in several cases. However, the model with $K = 4$ under the Train:Test ratio of 70:30 provides the best overall forecasting performance.

This model yields an RMSE of 96.557, an MAE of 70.935, and a MAPE of 3.423, which are the lowest forecasting errors compared with most of the other models. In addition, the Ljung–Box test indicates that the residuals do not exhibit autocorrelation, since the p-value is greater than 0.05.

Although the case with a Train:Test ratio of 80:20 and $K = 8$ yields the lowest MAPE value of 3.375, the overall forecasting errors in terms of RMSE and MAE remain higher than those of the model with $K = 4$ under the Train:Test ratio of 70:30. Therefore, the MIDAS-AR-HMM model with $K = 4$ under the Train:Test ratio of 70:30 is selected as the most appropriate model for representing the structure of the MIDAS-AR-HMM model in this study.

The general form of the MIDAS-AR-HMM model can be expressed as follows:

$$Y_t = \begin{cases} \beta_{0,1} + \phi_1 Y_{t-1} + \\ \beta_{1,1} \sum_{k=1}^{4} w_{1,k}(\nu_{1,1}) X_{t-k/m}^{DXY} + \beta_{2,1} \sum_{k=1}^{4} w_{2,k}(\nu_{2,1}) X_{t-k/m}^{OIL} + \varepsilon_t, & S_t = 1, \\ \beta_{0,2} + \phi_2 Y_{t-1} + \\ \beta_{1,2} \sum_{k=1}^{4} w_{1,k}(\nu_{1,2}) X_{t-k/m}^{DXY} + \beta_{2,2} \sum_{k=1}^{4} w_{2,k}(\nu_{2,2}) X_{t-k/m}^{OIL} + \varepsilon_t, & S_t = 2. \end{cases}$$

where S_t denotes the hidden regime at time t, $(S_t = 1, 2)$, determined by the Hidden Markov Model. The terms $w_{1,k}$ and $w_{2,k}$ denote the lag weights under the Exponential Almon Lag Polynomial for the DXY index and oil price variables, respectively. The estimated parameters of the MIDAS-AR-HMM model that provides the best forecasting performance are presented in Appendix A.4.

4.7 Results of the MIDAS-AR-HMM-SOM Model

4.7.1 Regime Classification Using the Self-Organizing Map and Hidden Markov Model

In this study, the MIDAS-AR-HMM-SOM model is developed from the MIDAS-AR-HMM model by incorporating the Self-Organizing Map (SOM) technique to cluster data characteristics before applying the Hidden Markov Model (HMM), with the aim of improving the classification of gold market regimes.

The variables used in the analysis consist of the logarithmic return of gold prices, volatility, and lagged return, which can be expressed as follows:

$$D_t = (r_t, \sigma_t, r_{t-1}),$$

where r_t denotes the logarithmic return of the gold price at time t, σ_t denotes volatility, which is calculated using the rolling standard deviation of logarithmic returns, and r_{t-1} denotes the lagged return of the gold price.

Before entering the SOM process, all data are standardized using the Standardization method to reduce the effect of differences in data scales. The standardized data are then processed by the Self-Organizing Map (SOM) to group observations with similar characteristics. The resulting clusters are subsequently used as inputs to the Gaussian Hidden Markov Model (HMM) for classifying the hidden regimes of the gold market.

In this study, the Gaussian Hidden Markov Model is specified with two hidden states to represent different market regimes. The HMM learns the probability distribution of the data under each regime while considering the temporal continuity of regimes. In addition, the HMM uses a Transition Probability Matrix to describe the probabilities of remaining in or switching between regimes over time, which can be expressed as follows:

$$P = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix},$$

where p_{ij} denotes the probability that the market transitions from regime i to regime j

in the next period.

For this study, the estimated Transition Probability Matrix is given by

$$P = \begin{pmatrix} 0.468 & 0.532 \\ 0.299 & 0.701 \end{pmatrix}.$$

4.7.2 Analysis of AIC and BIC Values for the MIDAS-AR-HMM-SOM Model

The information criteria results for the MIDAS-AR-HMM-SOM model are presented in Table 4.19.

Table 4.19 AIC and BIC Values of the MIDAS-AR-HMM-SOM Model.

Train:Test	$K = 4$		$K = 8$		$K = 12$	
	AIC	BIC	AIC	BIC	AIC	BIC
70:30	1515.225	1568.548	1509.521	1562.844	1512.470	1565.793
75:25	1622.531	1676.974	1617.403	1671.846	1620.268	1674.711
80:20	1737.543	1792.964	1737.445	1792.866	1740.144	1795.565

Table 4.19 shows that the MIDAS-AR-HMM-SOM model with $K = 8$ yields the lowest AIC and BIC values under all Train:Test ratios. This indicates that, based on the information criteria, the model with $K = 8$ is more appropriate than the other cases in terms of balancing goodness of fit and model complexity.

4.7.3 Residual Diagnostic Checking

Table 4.20 Ljung–Box Test Results for the MIDAS-AR-HMM-SOM Model.

Train:Test	$K = 4$	$K = 8$	$K = 12$
	Ljung–Box p-value	Ljung–Box p-value	Ljung–Box p-value
70:30	0.053	0.067	0.126
75:25	0.041	0.056	0.099
80:20	0.121	0.137	0.185

Table 4.20 shows that the Ljung–Box test results for the MIDAS-AR-HMM-SOM model generally produce p-values greater than 0.05, indicating that the residuals of most models do not exhibit statistically significant autocorrelation. In particular, the cases with

$K = 8$ and $K = 12$ satisfy the Ljung–Box test criterion under all Train:Test ratios. In contrast, the case with a Train:Test ratio of 75:25 and $K = 4$ yields a p-value of 0.041, which is slightly less than 0.05.

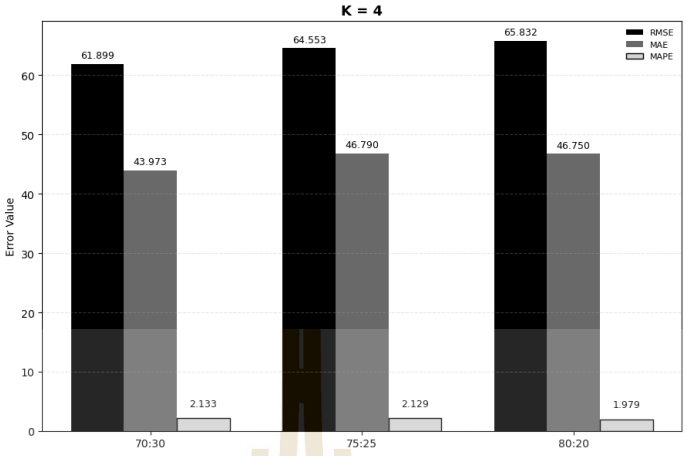
4.7.4 Performance Evaluation of the Model on the Test Dataset

Table 4.21 Performance Evaluation Results of the MIDAS-AR-HMM-SOM Model on the Test Dataset.

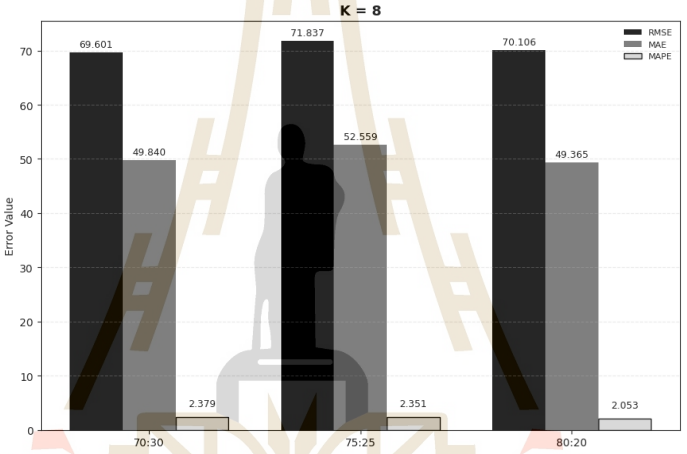
Train:Test	$K = 4$			$K = 8$			$K = 12$		
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	RMSE	MAE	MAPE
70:30	61.899	43.973	2.133	69.601	49.840	2.379	66.677	49.931	2.417
75:25	64.553	46.790	2.129	71.837	52.559	2.351	68.986	52.340	2.373
80:20	65.832	46.750	1.979	70.106	49.365	2.053	67.355	48.695	2.049

Table 4.21 shows that the MIDAS-AR-HMM-SOM model produces low forecasting errors in all cases, with MAPE values below 3% for all Train:Test ratios. Based on the forecasting error measures, the model with $K = 4$ yields the lowest RMSE and MAE values for the Train:Test ratios of 70:30 and 75:25. In contrast, for the Train:Test ratio of 80:20, the model with $K = 4$ produces the lowest MAPE value of 1.979.

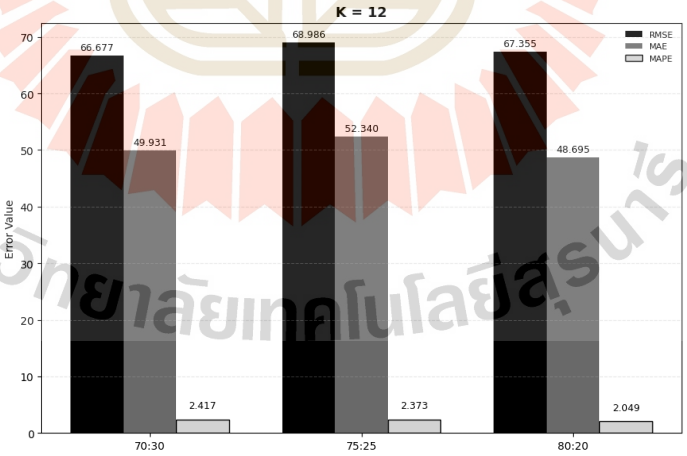
มหาวิทยาลัยเทคโนโลยีสุรนารี



(a) $K = 4$



(b) $K = 8$



(c) $K = 12$

Figure 4.6 Forecasting Performance of the MIDAS-AR-HMM-SOM Model under Different K Values and Train:Test Ratios.

4.7.5 Selection of the Most Appropriate MIDAS-AR-HMM-SOM Model

Based on the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Ljung–Box test results, and forecasting error measures, including RMSE, MAE, and MAPE, the MIDAS-AR-HMM-SOM model provides good better forecasting performance under all data partition schemes.

The AIC and BIC results indicate that the model with $K = 8$ yields the lowest information criteria values under all Train:Test ratios. However, when the forecasting performance on the test dataset is also considered, the model with $K = 4$ produces lower forecasting errors in several cases. In particular, for the Train:Test ratio of 70:30, the model with $K = 4$ yields an RMSE of 61.899, an MAE of 43.973, and a MAPE of 2.133, which are the lowest RMSE and MAE values among all models. For the Train:Test ratio of 80:20, the model with $K = 4$ produces the lowest MAPE value of 1.979.

In addition, the Ljung–Box test for the case with a Train:Test ratio of 70:30 and $K = 4$ yields a p-value of 0.053, which is greater than 0.05, indicating that the residuals do not exhibit autocorrelation.

Therefore, the MIDAS-AR-HMM-SOM model with $K = 4$ under the Train:Test ratio of 70:30 is selected as the most appropriate model in this study, since it provides the lowest overall forecasting errors and satisfies the residual diagnostic criterion.

The general form of the MIDAS-AR-HMM-SOM model can be expressed as follows:

$$Y_t = \begin{cases} \beta_{0,1} + \phi_1 Y_{t-1} + \\ \beta_{1,1} \sum_{k=1}^{4} w_{1,k}(\nu_{1,1}) X^{DXY}_{t-k/m} + \beta_{2,1} \sum_{k=1}^{4} w_{2,k}(\nu_{2,1}) X^{OIL}_{t-k/m} + \varepsilon_t, & S_t = 1, \\ \beta_{0,2} + \phi_2 Y_{t-1} + \\ \beta_{1,2} \sum_{k=1}^{4} w_{1,k}(\nu_{1,2}) X^{DXY}_{t-k/m} + \beta_{2,2} \sum_{k=1}^{4} w_{2,k}(\nu_{2,2}) X^{OIL}_{t-k/m} + \varepsilon_t, & S_t = 2. \end{cases}$$

where S_t denotes the hidden regime at time t, $(S_t = 1, 2)$, obtained from the Self-Organizing Map combined with the Hidden Markov Model. The terms $w_{1,k}$ and $w_{2,k}$ denote the lag weights under the Exponential Almon Lag Polynomial for the DXY index and oil price variables, respectively. The estimated parameters of the MIDAS-AR-HMM-

SOM model that provides the best forecasting performance are presented in Appendix A.5.

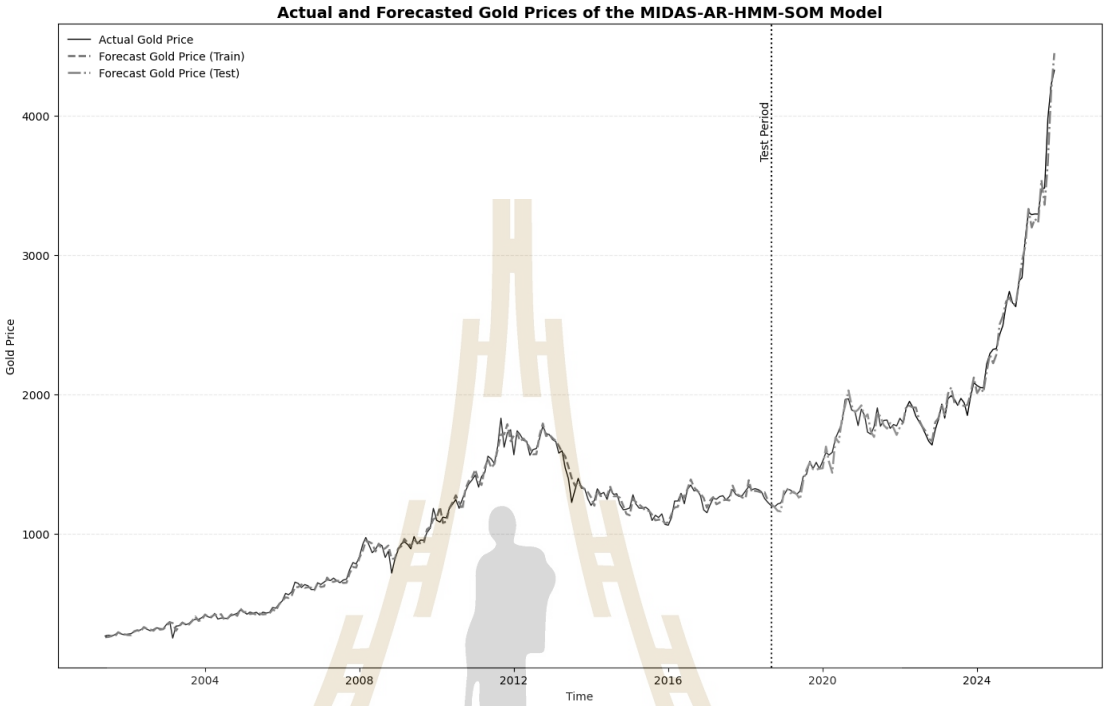


Figure 4.7 Actual and Forecasted Gold Prices of the MIDAS-AR-HMM-SOM Model during the Training and Testing Periods.

4.8 Comparison of Results under Different Train:Test Ratios

In this section, the forecasting performance of the models is compared under different Train:Test ratios, namely 80:20, 75:25, and 70:30. The Diebold–Mariano Test (DM Test) is employed to examine differences in forecasting accuracy between models. The test results are presented separately for each lag specification of the MIDAS models, namely $K = 4$, $K = 8$, and $K = 12$, to facilitate a clear comparison of model performance under different lag structures.

Note: A positive DM Statistic indicates that Model 1 provides more accurate forecasts than Model 2, whereas a negative DM Statistic indicates that Model 2 provides more accurate forecasts than Model 1. However, if the p-value is greater than 0.05, the difference is considered statistically insignificant.

Table 4.22 Diebold–Mariano Test Results for $K = 4$ (Train:Test = 70:30).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	4.553	0.000
MIDAS-AR-HMM-SOM	MIDAS	12.446	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	3.126	0.002
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	4.090	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	3.423	0.00

Table 4.23 Diebold–Mariano Test Results for $K = 8$ (Train:Test = 70:30).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	4.554	0.000
MIDAS-AR-HMM-SOM	MIDAS	12.450	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	2.581	0.011
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	3.747	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	2.557	0.012

Table 4.24 Diebold–Mariano Test Results for $K = 12$ (Train:Test = 70:30).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	4.549	0.00
MIDAS-AR-HMM-SOM	MIDAS	12.445	0.00
MIDAS-AR-HMM-SOM	MIDAS-AR	2.725	0.008
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	3.675	0.00
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	2.877	0.005

มหาวิทยาลัยเทคโนโลยีสุรนารี

Table 4.25 Diebold–Mariano Test Results for $K = 4$ (Train:Test = 75:25).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	4.004	0.000
MIDAS-AR-HMM-SOM	MIDAS	12.491	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	3.115	0.003
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	3.845	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	3.389	0.001

Table 4.26 Diebold–Mariano Test Results for $K = 8$ (Train:Test = 75:25).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	4.004	0.000
MIDAS-AR-HMM-SOM	MIDAS	12.495	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	2.698	0.009
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	3.183	0.002
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	2.612	0.010

Table 4.27 Diebold–Mariano Test Results for $K = 12$ (Train:Test = 75:25).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	4.003	0.000
MIDAS-AR-HMM-SOM	MIDAS	12.490	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	2.701	0.009
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	3.442	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	2.888	0.005

มหาวิทยาลัยเทคโนโลยีสุรนารี

Table 4.28 Diebold–Mariano Test Results for $K = 4$ (Train:Test = 80:20).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	3.353	0.001
MIDAS-AR-HMM-SOM	MIDAS	11.611	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	2.975	0.004
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	2.852	0.006
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	2.996	0.003

Table 4.29 Diebold–Mariano Test Results for $K = 8$ (Train:Test = 80:20).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	3.355	0.001
MIDAS-AR-HMM-SOM	MIDAS	11.614	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	2.660	0.010
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	2.700	0.009
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	2.385	0.020

Table 4.30 Diebold–Mariano Test Results for $K = 12$ (Train:Test = 80:20).

Model 1	Model 2	DM-Statistic	p-value
MIDAS-AR-HMM-SOM	ARIMA	3.353	0.001
MIDAS-AR-HMM-SOM	MIDAS	11.611	0.000
MIDAS-AR-HMM-SOM	MIDAS-AR	2.712	0.009
MIDAS-AR-HMM-SOM	MIDAS-AR-Threshold	2.734	0.008
MIDAS-AR-HMM-SOM	MIDAS-AR-HMM	2.568	0.013

มหาวิทยาลัยเทคโนโลยีสุรนารี

CHAPTER V

DISCUSSION AND CONCLUSION

5.1 Discussion

5.1.1 Discussion of the ARIMA Model

The ARIMA model was used as the baseline model for comparison with the other models investigated in this study. The model development was conducted under different training and testing data ratios, namely 70:30, 75:25, and 80:20. In constructing the ARIMA model, it was necessary to examine the stationarity of the data beforehand, since the ARIMA model can only be developed under the condition that the data are stationary (Weak Stationarity). Therefore, the researcher tested the stationarity of the data using the Augmented Dickey–Fuller (ADF) Test. The results indicated that the gold price data at the original level were non-stationary under all data-splitting ratios. Consequently, the first-order differencing was applied, and the data were found to become stationary after this transformation.

The results of the model selection analysis for all data-splitting ratios showed that ARIMA(0,1,1) was the most appropriate model according to the selection criteria employed in this study. The residual diagnostic results further indicated that the ARIMA(0,1,1) model did not exhibit any autocorrelation problem in the error terms in all cases. However, when considering the forecasting errors, the values remained relatively high compared with those of the models developed in this research. This finding suggests that the use of only historical gold price data may not be sufficient to explain the long-term behavior of gold prices, which are influenced by economic factors and complex market conditions. Furthermore, the results also demonstrated that although ARIMA can effectively handle the non-stationarity problem of the data, a major limitation of the model is its inability to incorporate external variables that may influence gold prices into the forecasting process.

5.1.2 Discussion of the MIDAS Model

The MIDAS model was employed in this study because of its ability to incorporate data with different frequencies within a single model. Monthly gold prices were specified as the dependent variable, while the DXY index and weekly oil prices were used as independent variables. Three lag structures were considered, namely $(K = 4)$, $(K = 8)$, and $(K = 12)$, under different training and testing data ratios.

The results of the AIC and BIC analyses indicated that the most appropriate model varied across different data-splitting ratios. For the Train:Test ratios of 70:30 and 75:25, the model with $(K = 12)$ yielded the lowest information criteria values, whereas for the Train:Test ratio of 80:20, the model with $(K = 4)$ produced the lowest information criteria values. This finding suggests that the optimal number of lags may vary according to the amount of data used for training. However, when considering the residual diagnostic results using the Ljung–Box Test, it was found that the residuals of the MIDAS model still exhibited autocorrelation problems in all cases. This indicates that although the model is capable of incorporating data with different frequencies, it is still unable to adequately explain the structure of gold price data. When considering forecasting performance, the MIDAS model produced relatively high forecasting errors in all cases and exhibited lower performance than the ARIMA(0,1,1) model. The RMSE values exceeded one thousand units for all data-splitting ratios, despite incorporating the DXY index and oil prices, which are factors related to gold prices, into the model.

These findings indicate that the inclusion of high-frequency data alone is not sufficient for forecasting gold prices. This is because gold prices are influenced not only by external factors but also by their own historical price movements. Therefore, although the MIDAS model can effectively address the problem of combining data with different frequencies, it still has limitations in explaining the temporal relationships within gold price data. This limitation is expected to be an important reason why its forecasting performance remains unsatisfactory.

5.1.3 Discussion of the MIDAS-AR Model

The MIDAS-AR model extends the MIDAS model by incorporating an autoregressive component into its structure in order to consider the effect of historical gold price data on forecasting. The results of the AIC and BIC analyses indicated that, for the Train:Test ratios of 70:30 and 75:25, the model with $(K = 4)$ produced the lowest information criteria values, whereas for the Train:Test ratio of 80:20, the model with $(K = 12)$ yielded the lowest AIC and BIC values. When considering the residual diagnostic results, it was found that the MIDAS-AR model passed the Ljung–Box test only in the case of Train:Test = 80:20 with $(K = 12)$, indicating no autocorrelation problem in the residuals. In contrast, the other cases still exhibited such a problem. Regarding forecasting performance, the MIDAS-AR model showed substantially lower forecasting errors compared with the MIDAS model. In particular, the case of Train:Test = 80:20 with $(K = 12)$ produced the lowest MAE and MAPE values. Under this data-splitting ratio, the MIDAS-AR model with $(K = 12)$ was selected as the most appropriate model.

These findings indicate that incorporating historical gold price data into the forecasting process can significantly improve model performance. This suggests that gold prices are influenced not only by the DXY index and oil prices but also by their own historical price movements.

5.1.4 Discussion of the MIDAS-AR-Threshold Model

The MIDAS-AR-Threshold model is an extension of the MIDAS-AR model that incorporates a regime-switching mechanism through the threshold value of the gold price lagged by one period, Y_{t-1}, in order to investigate the effect of changes in market conditions on forecasting performance.

The results of the AIC and BIC analyses indicated that the model with $(K = 12)$ produced the lowest information criteria values under all data-splitting ratios. This suggests that using a larger number of lags enables the model to better explain the data under a regime-switching structure. When considering the residual diagnostic results, it was found that the MIDAS-AR-Threshold model passed the Ljung–Box test in some cases,

particularly for $K = 4$ and $K = 8$ under the Train:Test ratios of 70:30 and 75:25, as well as for $K = 12$ under the Train:Test ratio of 80:20, where the p-values were greater than 0.05. This indicates that the residuals of the model in these cases did not exhibit statistically significant autocorrelation. However, some cases still produced p-values lower than 0.05, indicating that the MIDAS-AR-Threshold model did not pass the residual diagnostic test in all cases. Regarding forecasting performance, the MIDAS-AR-Threshold model produced relatively low forecasting errors, but overall the errors were slightly higher than those of the MIDAS-AR model. The most appropriate MIDAS-AR-Threshold model was the model with a Train:Test ratio of 80:20 and $(K = 12)$, as it yielded the lowest AIC and BIC values under the same data-splitting ratio. Furthermore, no autocorrelation problem was detected in the residuals, and the forecasting errors remained at a relatively low level. Although, under the same data-splitting ratio, the model with $(K = 4)$ produced slightly lower RMSE, MAE, and MAPE values, that model still suffered from an autocorrelation problem.

These findings indicate that dividing the data into regimes based on the level of gold prices enables the model to account for differences in data behavior under different market conditions. However, defining regimes using a single threshold value remains a fixed boundary approach, which may not fully capture actual changes in market conditions. Furthermore, when considering forecasting errors, the MIDAS-AR-Threshold model still produced higher RMSE, MAE, and MAPE values than the MIDAS-AR model in several experimental cases. This suggests that regime classification based solely on a threshold value may not be sufficient to explain complex market dynamics. Therefore, the application of a Hidden Markov Model for regime classification was investigated in the subsequent stage of the study.

5.1.5 Discussion of the MIDAS-AR-HMM Model

The MIDAS-AR-HMM model is an extension of the MIDAS-AR-Threshold model that employs a Hidden Markov Model (HMM) to classify market regimes instead of using a fixed threshold. The market regime classification was based on gold returns and the volatility of gold returns as observable variables, since these two variables can better reflect changes

in the gold market than the gold price level alone. Gold returns represent price changes over time, while volatility reflects the level of market fluctuations. Considering these two variables simultaneously enables the Hidden Markov Model to distinguish different market conditions more clearly, which is consistent with the concept of regime-switching that aims to explain changes in market behavior over time.

The results of the AIC and BIC analyses indicated that the model with $(K = 12)$ produced the lowest information criteria values under all data-splitting ratios. This finding suggests that, within the MIDAS-AR-HMM framework, $(K = 12)$ provides a better model fit according to the information criteria than $K = 4$ and $K = 8$. When considering the residual diagnostic results, it was found that the MIDAS-AR-HMM model passed the Ljung–Box test in almost all experimental cases. Autocorrelation was detected only in the case of Train:Test = 75:25 with $(K = 4)$. This result indicates that the model was able to explain the underlying data structure more effectively than the MIDAS-AR-Threshold model. Regarding forecasting performance, the MIDAS-AR-HMM model produced lower forecasting errors than the MIDAS-AR-Threshold model. The model with a Train:Test ratio of 70:30 and $(K = 4)$ provided the best overall forecasting performance, as it yielded the lowest RMSE, MAE, and MAPE values while exhibiting no autocorrelation problem in the residuals. Therefore, it was selected as the most appropriate MIDAS-AR-HMM model in this study.

These findings indicate that the Hidden Markov Model can classify market regimes more flexibly than a fixed-threshold approach. This is because the HMM can learn changes in market conditions directly from the data and takes into account the temporal continuity of regimes. Furthermore, the Transition Probability Matrix showed that each regime exhibited a relatively high degree of persistence over time, which is consistent with the behavior of the gold market, where market conditions tend to remain in the same state for a period before transitioning to a new state. However, although the MIDAS-AR-HMM model provided improved forecasting performance, regime classification was still based solely on the Hidden Markov Model. Therefore, the application of a Self-Organizing Map, which is a machine learning technique used for data clustering, in combination with the Hidden Markov Model may provide a more refined regime classification and may further

improve the forecasting performance of the model.

5.1.6 Discussion of the MIDAS-AR-HMM-SOM Model

The MIDAS-AR-HMM-SOM model is an extension of the MIDAS-AR-HMM model that integrates the Self-Organizing Map (SOM) technique with the Hidden Markov Model to assist in clustering the data prior to the classification of gold market regimes. In this study, logarithmic returns, volatility, and lagged returns of gold prices were used as variables for analysis because these three variables can reflect price changes, the level of volatility, and the influence of historical information, all of which are important factors affecting changes in market conditions.

The results of the AIC and BIC analyses indicated that the model with $(K = 8)$ produced the lowest information criteria values under all data-splitting ratios. This finding suggests that the model provides the best ability to explain the data when compared with the other cases considered in this study. When considering the residual diagnostic results, it was found that the MIDAS-AR-HMM-SOM model passed the Ljung–Box test in almost all experimental cases. In particular, the cases of $(K = 8)$ and $(K = 12)$ passed the test under all data-splitting ratios, whereas the case of Train:Test = 75:25 with $(K = 4)$ was the only case that produced a p-value lower than 0.05. Regarding forecasting performance, the MIDAS-AR-HMM-SOM model produced low forecasting errors in all experimental cases. The MAPE values were below 3% for all data-splitting ratios. Furthermore, when compared with the MIDAS-AR-HMM model, the forecasting errors were substantially reduced. In particular, the case of Train:Test = 70:30 with $(K = 4)$ yielded an RMSE of 61.90, an MAE of 43.97, and a MAPE of 2.13, while also passing the Ljung–Box test. Therefore, this model was selected as the most appropriate model in this study.

The findings indicate that integrating SOM with HMM enables a more refined regime classification. This is because SOM can cluster data with similar characteristics before the learning process of the HMM, resulting in lower forecasting errors compared with the MIDAS-AR-HMM model. This improvement suggests that gold market data may possess a more complex structure than can be captured by regime classification using HMM alone, and that the inclusion of a data-clustering stage prior to regime classification helps the

model learn the underlying characteristics of the data more effectively. When considered together with the results of all models employed in this research, the MIDAS-AR-HMM-SOM model provided the best forecasting performance in this study, both in terms of forecasting accuracy and residual diagnostic results. This demonstrates that the integration of mixed-frequency data, the consideration of the influence of historical information, and regime classification through machine learning techniques can effectively enhance the forecasting performance of gold prices.

5.2 Conclusion

This study demonstrates that accurate forecasting of mixed-frequency financial time series requires more than the incorporation of high-frequency explanatory variables. The empirical findings indicate that forecasting performance depends critically on the ability of a model to account for both the temporal dependence embedded in historical observations and the dynamic changes in underlying market conditions. Consequently, the effectiveness of a forecasting framework is determined not only by the quantity or frequency of available information but also by its capacity to represent the structural characteristics of the data-generating process.

The results consistently show that regime-dependent modeling provides substantial benefits for forecasting gold prices. Financial markets are inherently characterized by periods of expansion, contraction, stability, and heightened uncertainty, and the relationships between explanatory variables and gold prices may vary across these market environments. Models that explicitly incorporate regime information were therefore able to capture changes in market behavior more effectively than models based solely on a single, fixed structure. In particular, probabilistic regime identification through the Hidden Markov Model proved more effective than deterministic threshold-based classification, suggesting that latent-state representations are better suited to describing the complex and evolving nature of financial markets.

A further contribution of this research is the integration of Self-Organizing Maps with the Hidden Markov Model within the MIDAS framework. The findings indicate that

combining unsupervised learning techniques with regime-switching methodologies enhances the identification of market states and improves the model's ability to distinguish between heterogeneous patterns in mixed-frequency data. This hybrid approach enables a more flexible representation of market dynamics and leads to superior forecasting performance across different experimental settings.

From a methodological perspective, the study extends the application of MIDAS models by incorporating autoregressive structures, latent regime-switching mechanisms, and machine-learning-based state classification into a unified forecasting framework. The superior performance of the MIDAS-AR-HMM-SOM model, supported by both forecasting accuracy measures and Diebold–Mariano statistical tests, demonstrates the value of integrating econometric and machine learning approaches for mixed-frequency financial forecasting.

Overall, the findings suggest that the combination of mixed-frequency information, temporal dependence, and adaptive regime identification provides a robust framework for forecasting gold prices under changing market conditions. The proposed MIDAS-AR-HMM-SOM model therefore offers a meaningful methodological contribution to the literature on mixed-frequency time series forecasting and provides a practical approach for improving predictive performance in complex financial environments.

มหาวิทยาลัยเทคโนโลยีสุรนารี

REFERENCES

มหาวิทยาลัยเทคโนโลยีสุรนารี

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Andreou, E., Ghysels, E., and Kourtellos, A. (2010). Regression models with mixed sampling frequencies. *Journal of Econometrics*, *158*(2), 246–261.

Atsalakis, G. S., and Valavanis, K. P. (2009). Surveying stock market forecasting techniques – Part II: Soft computing methods. *Expert Systems with Applications*, *36*(3), 5932–5941.

Bahelka, A., and de Weerd, H. (2024). Comparative analysis of Mixed-Data Sampling (MIDAS) model compared to Lag-Llama model for inflation nowcasting. *arXiv*, arXiv:2407.08510.

Barsoum, F., and Stankiewicz, S. (2015). Forecasting GDP growth using mixed-frequency models with switching regimes. *Working Paper*, University of Konstanz.

Baum, L. E., and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, *37*(6), 1554–1563.

Baur, D. G., and McDermott, T. K. (2010). Is gold a safe haven? International evidence. *Journal of Banking & Finance*, *34*(8), 1886–1898. https://doi.org/10.1016/j.jbankfin.2009.12.008

Beckmann, J., Berger, T., and Czudaj, R. (2015). Does gold act as a hedge or a safe haven for stocks? A smooth transition approach. *Economic Modelling*, *48*, 16–24. https://doi.org/10.1016/j.econmod.2014.10.044

Box, G. E. P., and Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control* (5th ed.). Wiley.

มหาวิทยาลัยเทคโนโลยีสุรนารี

Brockwell, P. J., and Davis, R. A. (2016). *Introduction to Time Series and Forecasting* (3rd ed.). Springer.

Burnham, K. P., and Anderson, D. R. (2002). *Model Selection and Multimodel Inference* (2nd ed.). Springer.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Cavicchioli, M. (2021). Markov Switching GARCH Models: Higher order moments, kurtosis measures, and volatility evaluation in recessions and pandemic. *Journal of Business & Economic Statistics*, *40*(4), 1772–1783. https://doi.org/10.1080/07350015.2021.1974459

Chan, J. C. C., Poon, A., and Zhu, D. (2024). Time-varying parameter MIDAS models: Application to nowcasting US real GDP. *Working Paper*.

Clements, M. P., and Galvão, A. B. (2008). Macroeconomic forecasting with mixed-frequency data: Forecasting output growth in the United States. *Journal of Business & Economic Statistics*, *26*(4), 546–554.

Cryer, J. D., and Chan, K.-S. (2008). *Time Series Analysis with Applications in R* (2nd ed.). Springer.

Dickey, D. A., and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366), 427–431.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). Wiley.

Engle, R. F., Ghysels, E., and Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*, *95*(3), 776–797.

มหาวิทยาลัยเทคโนโลยีสุรนารี

Ersin, Ö. Ö., and Bildirici, M. (2023). Financial volatility modeling with the GARCH-MIDAS-LSTM approach: The effects of economic expectations, geopolitical risks and industrial production during COVID-19. *Mathematics*, *11*(1785). https://doi.org/10.3390/math11081785

Fang, T., Lee, T.-H., and Su, Z. (2020). Predicting the long-term stock market volatility: A GARCH-MIDAS model with variable selection. *Journal of Empirical Finance*, *58*, 36–49. https://doi.org/10.1016/j.jempfin.2020.05.007

Foroni, C., and Marcellino, M. (2013). A survey of econometric methods for mixed-frequency data. *Norges Bank Working Paper*, No. 2013/06.

Foroni, C., Marcellino, M., and Schumacher, C. (2015). U-MIDAS: MIDAS regressions with unrestricted lag polynomials. *Journal of the Royal Statistical Society: Series A*, *178*(1), 57–82.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS touch: Mixed Data Sampling regression models. University of North Carolina and UCLA.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2006). Predicting volatility: Getting the most out of return data sampled at different frequencies. *Journal of Econometrics*, *131*(1–2), 59–95. https://doi.org/10.1016/j.jeconom.2005.01.004

Ghysels, E., Sinko, A., and Valkanov, R. (2007). MIDAS regressions: Further results and new directions. *Econometric Reviews*, *26*(1), 53–90.

Ghysels, E., and Qian, H. (2019). Estimating MIDAS regressions via OLS with polynomial parameter profiling. *Econometrics and Statistics*, *9*, 1–16. https://doi.org/10.1016/j.ecosta.2018.02.001

Giovanis, E. (2010). Application of logit model and self-organizing maps (SOMs) for the prediction of financial crisis periods in US economy. *Journal of Financial Economic Policy*, *2*(2), 98–125. https://doi.org/10.1108/17576381011070184

Granger, C. W. J., and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics*, *2*(2), 111–120.

Guérin, P., and Marcellino, M. (2013). Markov-switching MIDAS models. *Journal of Business & Economic Statistics*, *31*(1), 45–56.

Gujarati, D. N., and Porter, D. C. (2009). *Basic Econometrics* (5th ed.). McGraw-Hill.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*(2), 357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Hamilton, J. D., and Susmel, R. (1994). Autoregressive conditional heteroskedasticity and changes in regime. *Journal of Econometrics*, *64*(1–2), 307–333. https://doi.org/10.1016/0304-4076(94)90067-1

Harvey, D., Leybourne, S., and Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291.

Hyndman, R. J., and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.).

Kiviluoto, K. (1998). Predicting bankruptcies with the self-organizing map. *Neurocomputing*, *21*(1–3), 191–201. https://doi.org/10.1016/S0925-2312(98)00038-1

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Kohonen, T. (2001). *Self-Organizing Maps* (3rd ed.). Springer.

Kole, E., and van Dijk, D. (2023). Moments, shocks and spillovers in Markov-switching VAR models. *Journal of Econometrics*, *236*(2), 105474. https://doi.org/10.1016/j.jeconom.2023.105474

Ljung, G. M., and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, *65*(2), 297–303.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Ma, W., Xu, Y., Xu, Z., and Zhang, Y. (2024). On stock volatility forecasting under mixed-frequency data. *Mathematics*, *12*(10), 1538. https://doi.org/10.3390/math12101538

Marcellino, M., and Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: A model comparison for German GDP. *Oxford Bulletin of Economics and Statistics*, *72*(4), 518–550.

Nguyen, D.-H., and Le, M.-T. (2014). A two-stage architecture for stock price forecasting by combining SOM and fuzzy-SVM. *arXiv preprint arXiv:1408.5241*.

Nocedal, J., and Wright, S. J. (2006). *Numerical Optimization* (2nd ed.). Springer.

Said, S. E., and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, *71*(3), 599–607.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Slutsky, E. (1927). The summation of random causes as the source of cyclic processes. *Econometrica*, *5*(2), 105–146.

Tan, C.-Y., Koh, Y.-B., Ng, K.-H., and Ng, K.-H. (2021). Dynamic volatility modelling of Bitcoin using time-varying transition probability Markov-switching GARCH model. *The North American Journal of Economics and Finance*, *56*, 101377. https://doi.org/10.1016/j.najef.2021.101377

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, *89*(425), 208–218. https://doi.org/10.2307/2291217

Tong, H. (1978). On a threshold model. In *Pattern Recognition and Signal Processing* (pp. 575–586).

Tong, H., and Lim, K. S. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society: Series B*, *42*(3), 245–292.

Tsay, R. S. (2010). *Analysis of Financial Time Series* (3rd ed.). Wiley.

Tsui, A. K. C., Xu, C. Y., and Zhang, Z. Y. (2013, December). Forecasting Singapore economic growth with mixed-frequency data. In *20th International Congress on Modelling and Simulation* (pp. 1399–1405). Adelaide, Australia: Modelling and Simulation Society of Australia and New Zealand. `https://www.mssanz.org.au/modsim2013/F8/tsui2.pdf`

Virk, N., Javed, F., and Awartani, B. (2021). A reality check on the GARCH-MIDAS volatility models. *Örebro University Working Paper*, No. 2/2021.

Wang, C., Xu, M., Wang, Z., and Sun, W. (2024). Research on China insurance demand forecasting: Based on mixed frequency data model. *PLOS ONE*, *19*(7), e0305523. `https://doi.org/10.1371/journal.pone.0305523`

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society A*, *226*, 267–298.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). L-BFGS-B: Algorithm 778. *ACM Transactions on Mathematical Software*, *23*(4), 550–560.

APPENDICES

มหาวิทยาลัยเทคโนโลยีสุรนารี

APPENDIX A

PARAMETER SETTINGS

มหาวิทยาลัยเทคโนโลยีสุรนารี

A.1 ESTIMATED PARAMETERS OF THE MIDAS MODEL

Table A.1 Estimated Parameters of the MIDAS Model

Parameter	Estimated Value
β_0	128.349
β_1	2.954
β_2	9.910
$\nu_1 = (\nu_{11}, \nu_{12})$	(5.117, 35.427)
$\nu_2 = (\nu_{21}, \nu_{22})$	(8.947, 41.842)

Note: ν_{11} and ν_{12} denote the first and second parameters of the Exponential Almon lag polynomial for the DXY index variable, respectively. Similarly, ν_{21} and ν_{22} denote the first and second parameters of the Exponential Almon lag polynomial for the OIL variable, respectively.

Table A.2 MIDAS Lag Weights Based on the Estimated Exponential Almon Polynomial

k	DXY Weight $w_{1,k}(\nu_1)$	OIL Weight $w_{2,k}(\nu_2)$
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000
4	1.000	1.000

A.2 ESTIMATED PARAMETERS OF THE MIDAS-AR MODEL

Table A.3 Estimated Parameters of MIDAS-AR Model

Parameter	Estimated Value
β_0	39.043
ϕ_1	0.999
β_1	-0.202
β_2	-0.214
$\nu_1 = (\nu_{11}, \nu_{12})$	(-0.033, 1.193)
$\nu_2 = (\nu_{21}, \nu_{22})$	(2.603, 1.224)

Note: ν_{11} denotes the first Exponential Almon parameter for the DXY index variable.

ν_{12} denotes the second Exponential Almon parameter for the DXY index variable.

ν_{21} denotes the first Exponential Almon parameter for the OIL price variable.

ν_{22} denotes the second Exponential Almon parameter for the OIL price variable.

มหาวิทยาลัยเทคโนโลยีสุรนารี

Table A.4 MIDAS-AR Lag Weights Based on the Estimated Exponential Almon Polynomial

k	DXY weight $w_{1,k}(\nu_1)$	OIL weight $w_{2,k}(\nu_2)$
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000
4	0.000	0.000
5	0.000	0.000
6	0.000	0.000
7	0.000	0.000
8	0.000	0.000
9	0.000	0.000
10	0.000	0.000
11	0.000	0.000
12	1.000	1.000

A.3 ESTIMATED PARAMETERS OF THE MIDAS-AR-THRESHOLD MODEL

Table A.5 Estimated Parameters of the MIDAS-AR-Threshold Model for Regime 1

Parameter	Estimated Value
$\beta_{0,1}$	0.268
ϕ_1	1.109
$\beta_{1,1}$	0.035
$\beta_{2,1}$	-0.093
$\nu_{1,1} = (\nu_{11}, \nu_{12})$	(4.007, -0.230)
$\nu_{2,1} = (\nu_{21}, \nu_{22})$	(0.363, 5.785)

where $\nu_{1,1}$ denotes the vector of Exponential Almon parameters for the DXY variable in Regime 1.

$\nu_{2,1}$ denotes the vector of Exponential Almon parameters for the OIL variable in Regime 1.

Table A.6 Estimated Parameters of the MIDAS-AR-Threshold Model for Regime 2

Parameter	Estimated Value
$\beta_{0,2}$	0.869
ϕ_2	0.984
$\beta_{1,2}$	0.434
$\beta_{2,2}$	-0.183
$\nu_{1,2} = (\nu_{11}, \nu_{12})$	(1.040, 0.002)
$\nu_{2,2} = (\nu_{21}, \nu_{22})$	(1.867, -0.79)

where $\nu_{1,2}$ denotes the vector of Exponential Almon parameters for the DXY variable in Regime 2.

$\nu_{2,2}$ denotes the vector of Exponential Almon parameters for the OIL variable in Regime 2.

Table A.7 MIDAS-AR-Threshold Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 1

k	DXY weight $w_{1,k}(\nu_{1,1})$	OIL weight $w_{2,k}(\nu_{2,1})$
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000
4	0.000	0.000
5	0.002	0.000
6	0.012	0.000
7	0.052	0.000
8	0.143	0.000
9	0.248	0.000
10	0.273	0.000
11	0.189	0.000
12	0.083	1.000

Table A.8 MIDAS-AR-Threshold Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 2

k	DXY weight $w_{1,k}(\nu_{1,2})$	OIL weight $w_{2,k}(\nu_{2,2})$
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000
4	0.000	0.001
5	0.000	0.003
6	0.001	0.009
7	0.003	0.025
8	0.009	0.057
9	0.026	0.112
10	0.076	0.189
11	0.224	0.272
12	0.660	0.334

มหาวิทยาลัยเทคโนโลยีสุรนารี

A.4 ESTIMATED PARAMETERS OF THE MIDAS-AR-HMM MODEL

Table A.9 Estimated Parameters of the MIDAS-AR-HMM Model for Regime 1

Parameter	Estimated Value
$\beta_{0,1}$	0.008
ϕ_1	0.023
$\beta_{1,1}$	-0.077
$\beta_{2,1}$	0.254
$\nu_{1,1} = (\nu_{11}, \nu_{12})$	(-0.610, -0.040)
$\nu_{2,1} = (\nu_{21}, \nu_{22})$	(2.010, -1.2085)

where $\nu_{1,1}$ denotes the vector of Exponential Almon parameters for the DXY variable in Regime 1.

$\nu_{2,1}$ denotes the vector of Exponential Almon parameters for the OIL variable in Regime 1.

Table A.10 Estimated Parameters of the MIDAS-AR-HMM Model for Regime 2

Parameter	Estimated Value
$\beta_{0,2}$	-0.010
ϕ_2	-0.491
$\beta_{1,2}$	-6.124
$\beta_{2,2}$	-1.183
$\nu_{1,2} = (\nu_{11}, \nu_{12})$	(0.431, -0.223)
$\nu_{2,2} = (\nu_{21}, \nu_{22})$	(-0.657, 0.129)

where $\nu_{1,2}$ denotes the vector of Exponential Almon parameters for the DXY variable in Regime 2.

$\nu_{2,2}$ denotes the vector of Exponential Almon parameters for the OIL variable in Regime 2.

Table A.11 MIDAS-AR-HMM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 1

k	DXY weight $w_{1,k}(\nu_{1,1})$	OIL weight $w_{2,k}(\nu_{2,1})$
1	0.530	0.272
2	0.277	0.606
3	0.133	0.120
4	0.059	0.002

Table A.12 MIDAS-AR-HMM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 2

k	DXY weight $w_{1,k}(\nu_{1,2})$	OIL weight $w_{2,k}(\nu_{2,2})$
1	0.271	0.411
2	0.333	0.238
3	0.263	0.178
4	0.133	0.172

A.5 ESTIMATED PARAMETERS OF THE MIDAS-AR-HMM-SOM MODEL

Table A.13 Estimated Parameters of the MIDAS-AR-HMM-SOM Model for Regime 1

Parameter	Estimated Value
$\beta_{0,1}$	0.060
ϕ_1	-0.288
$\beta_{1,1}$	-1.619
$\beta_{2,1}$	-0.392
$\nu_{1,1} = (\nu_{11}, \nu_{12})$	(0.149, 0.000)
$\nu_{2,1} = (\nu_{21}, \nu_{22})$	(0.145, 0.000)

where $\nu_{1,1}$ denotes the vector of Exponential Almon parameters for the DXY variable in Regime 1.

$\nu_{2,1}$ denotes the vector of Exponential Almon parameters for the OIL variable in Regime 1.

Table A.14 Estimated Parameters of the MIDAS-AR-HMM-SOM Model for Regime 2

Parameter	Estimated Value
$\beta_{0,2}$	-0.018
ϕ_2	-0.330
$\beta_{1,2}$	-0.047
$\beta_{2,2}$	0.375
$\nu_{1,2} = (\nu_{11}, \nu_{12})$	(0.002, -0.314)
$\nu_{2,2} = (\nu_{21}, \nu_{22})$	(0.102, -0.039)

where $\nu_{1,2}$ denotes the vector of Exponential Almon parameters for the DXY variable in Regime 2.

$\nu_{2,2}$ denotes the vector of Exponential Almon parameters for the OIL variable in Regime 2.

มหาวิทยาลัยเทคโนโลยีสุรนารี

Table A.15 MIDAS-AR-HMM-SOM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 1

k	DXY weight $w_{1,k}(\nu_{1,1})$	OIL weight $w_{2,k}(\nu_{2,1})$
1	0.197	0.209
2	0.229	0.234
3	0.266	0.263
4	0.308	0.295

Table A.16 MIDAS-AR-HMM-SOM Lag Weights Based on the Estimated Exponential Almon Polynomial for Regime 2

k	DXY weight $w_{1,k}(\nu_{1,2})$	OIL weight $w_{2,k}(\nu_{2,2})$
1	0.277	0.246
2	0.269	0.262
3	0.245	0.258
4	0.210	0.235

มหาวิทยาลัยเทคโนโลยีสุรนารี

APPENDIX B

PYTHON CODE

มหาวิทยาลัยเทคโนโลยีสุรนารี

This chapter presents the Python code developed for the MIDAS-AR-HMM-SOM model proposed in this thesis. The implementation provides the computational framework for model construction, parameter estimation, regime identification, and forecasting analysis.

```
# Import required libraries
import time
import numpy as np
import pandas as pd
from hmmlearn.hmm import GaussianHMM
from scipy.optimize import minimize
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import StandardScaler
from statsmodels.stats.diagnostic import acorr_ljungbox

# Set random seed
SEED = 4
np.random.seed(SEED)

# Input data
GOLD = [input data gold]
DXY = [input data DXY]
OIL = [input data OIL]

# Create date indices
gold_dates = pd.date_range("2001-01-31", periods=len(GOLD), freq="M")
weekly_dates = pd.date_range("2001-01-05", periods=len(DXY), freq="W-FRI")

gold_s = pd.Series(GOLD, index=gold_dates, dtype=float)
dxy_s = pd.Series(DXY, index=weekly_dates, dtype=float)
oil_s = pd.Series(OIL, index=weekly_dates, dtype=float)

# Compute logarithmic returns
gold_ret = np.log(gold_s).diff().dropna()
dxy_ret = np.log(dxy_s).diff().dropna()
oil_ret = np.log(oil_s).diff().dropna()
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
32
33 # Construct HMM-SOM features
34 vol = gold_ret.rolling(3).std()
35 lag_ret = gold_ret.shift(1)
36
37 df_feat = pd.DataFrame(
38     {
39         "ret": gold_ret,
40         "vol": vol,
41         "lag_ret": lag_ret,
42     }
43 ).dropna()
44
45 # Define Self-Organizing Map
46 class SOM:
47
48     def __init__(self, k, dim, seed):
49         np.random.seed(seed)
50         self.w = np.random.randn(k, dim)
51
52     def bmu(self, x):
53         return np.argmin(np.linalg.norm(self.w - x, axis=1))
54
55     def fit(self, X):
56         for t in range(1000):
57             x = X[t % len(X)]
58             b = self.bmu(x)
59             self.w[b] += 0.1 * (x - self.w[b])
60         return self
61
62     def transform(self, X):
63         return np.array([self.bmu(x) for x in X])
64
65 # Standardize features and apply SOM
66 scaler = StandardScaler()
67 X = scaler.fit_transform(df_feat[["ret", "vol", "lag_ret"]])
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
68
69  som = SOM(4, X.shape[1], SEED).fit(X)
70  df_feat["som"] = som.transform(X)
71
72  # Fit Hidden Markov Model
73  hmm_data = df_feat[["ret", "vol", "lag_ret", "som"]].values
74
75 hmm = GaussianHMM(
76         n_components=2,
77         n_iter=200,
78         random_state=SEED,
79  )
80
81 hmm.fit(hmm_data)
82  df_feat["regime"] = hmm.predict(hmm_data)
83
84
85  # Stable Exponential Almon weighting function
86  def almon(K, t1, t2):
87
88      j = np.arange(K)
89      z = t1 * j + t2 * j * j
90      z = z - np.max(z)
91
92      w = np.exp(z)
93      s = np.sum(w)
94
95      if (s == 0) or np.isnan(s) or np.isnan(w).any() or np.isinf(w).any():
96          return np.ones(K) / K
97
98      return w / s
99
100 # Extract last K weekly observations
101 def get_last_k(series, date, K):
102
103     vals = series.loc[series.index <= date].tail(K).values
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
104
105         if len(vals) == K:
106             return vals[::-1]
107
108         return None
109
110
111     # Build MIDAS dataset
112     K = 4
113     rows = []
114
115     idx = df_feat.index
116
117     for t in range(1, len(idx)):
118
119         dt = idx[t]
120
121         dxy_l = get_last_k(dxy_ret, dt, K)
122         oil_l = get_last_k(oil_ret, dt, K)
123
124         if dxy_l is None or oil_l is None:
125             continue
126
127         row = {
128             "date": dt,
129             "y": df_feat.loc[dt, "ret"],
130             "yl": df_feat.loc[idx[t - 1], "ret"],
131             "regime": df_feat.loc[dt, "regime"],
132         }
133
134         for j in range(K):
135             row[f"d{j}"] = dxy_l[j]
136             row[f"o{j}"] = oil_l[j]
137
138         rows.append(row)
139
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
df = pd.DataFrame(rows).set_index("date")

# Train-test split
TRAIN_RATIO = 0.70

split_idx = int(len(df) * TRAIN_RATIO)

train = df.iloc[:split_idx]
test = df.iloc[split_idx:]

# MIDAS-AR prediction function
def predict(p, Xd, Xo, yl):

    a, p1, bd, t1, t2, bo, s1, s2 = p

    wd = almon(K, t1, t2)
    wo = almon(K, s1, s2)

    y = a + p1 * yl + bd * (Xd @ wd) + bo * (Xo @ wo)

    if np.isnan(y).any():
        return np.zeros_like(yl)

    return y

# Objective function
def loss(p, Xd, Xo, yl, y):

    yhat = predict(p, Xd, Xo, yl)

    penalty = 0.01 * (
        p[3] ** 2
        + p[4] ** 2
        + p[6] ** 2
        + p[7] ** 2
    )
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
176
177         return np.sum((y - yhat) ** 2) + penalty
178
179 # Estimate MIDAS-AR parameters
180 def fit(df_sub):
181
182         Xd = df_sub[[f"d{i}" for i in range(K)]].values
183         Xo = df_sub[[f"o{i}" for i in range(K)]].values
184
185         y = df_sub["y"].values
186         yl = df_sub["yl"].values
187
188         init = [0, 0.2, 0.1, -0.1, -0.01, 0.1, -0.1, -0.01]
189
190         bounds = [
191             (-1, 1),
192             (-1, 1),
193             (-5, 5),
194             (-1, 1),
195             (-1, 0),
196             (-5, 5),
197             (-1, 1),
198             (-1, 0),
199         ]
200
201         res = minimize(
202             loss,
203             init,
204             args=(Xd, Xo, yl, y),
205             bounds=bounds,
206             method="L-BFGS-B",
207         )
208
209         return res.x
210
211 p0 = fit(train[train.regime == 0])
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
p1 = fit(train[train.regime == 1])

# Remaining functions (predict_df, return_to_price, calc_metrics,
# calc_aic_bic, and the print section) should follow the same
# indentation style shown above.
```


มหาวิทยาลัยเทคโนโลยีสุรนารี

CURRICULUM VITAE

NAME : Phakkhaphon Sawatkamon **GENDER :** Male

EDUCATION BACKGROUND:

- Master of Science (Applied Mathematics), Suranaree University of Technology, Thailand, 2020

SCHOLARSHIP:

- Kitti Bundit Suranaree University of Technology Scholarship

PUBLICATIONS AND CONFERENCE PROCEEDINGS:

- Sawatkamon, P., Thongkunwong, S., & Tanthanuch, J. (2025). Forecasting viral hepatitis patient numbers using Fourier series-modified GM(1,1) and DGM(1,1) models. In *Proceedings of the National and International Conference 2025: The 4th SUTH Conference 2025 Hospital* (pp. 199–208). Suranaree University of Technology.
- Sawaskamol, P., Kongmuangpak, P., Tanthanuch, J., & Rodjanadid, B. (2021). Analysis of the exchange rate of the Thai baht against the Chinese yuan using a support vector machine and firefly algorithm. *Suranaree Journal of Social Science*, *15*(1), 105–112.
- Sawaskamol, P., Rodjanadid, B., & Tanthanuch, J. (2020, May 23). Forecasting the exchange rate of Thai baht against the Chinese yuan using a support vector machine. In *Proceedings of the 5th National Conference on Quality Management and Innovation* (pp. 54–64), *Eastern University of Management and Technology*, Ubon Ratchathani, Thailand.

มหาวิทยาลัยเทคโนโลยีสุรนารี