

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.