MIDAS Regression for Forecasting GDP

1 Introduction

Time series play a crucial role in various fields, including statistics, economics, quantitative finance, and engineering. Their analysis is fundamental for applications such as economic forecasting, budgetary planning, stock market analysis, process control, and inventory management. A time series is typically defined as a sequence of observations recorded at equally spaced time intervals. The primary objectives of time series analysis include forecasting, anomaly detection, clustering, classification, and content-based querying.

Traditionally, researchers rely on classical linear models for time series analysis, assuming that all variables are sampled at the same frequency. However, in many real-world scenarios, particularly in macroeconomics, key indicators are not always observed at uniform intervals. Standard forecasting models struggle with this irregularity, making it challenging to integrate mixed-frequency data effectively. Mixed Data Sampling (MIDAS) provides a solution to this issue. Introduced in 2004 by Eric Ghysels, Pedro Santa-Clara, and Rossen Valkanov, MIDAS is an econometric regression and filtering technique that allows for independent variables to be sampled more frequently than the dependent variable. While initially developed for volatility forecasting, MIDAS has since demonstrated its utility in macroeconomic modeling.

Recent studies have leveraged MIDAS models to enhance forecasting accuracy. For example, Clements and Galvão (2008) and Marcellino and Schumacher (2010) applied MIDAS to predict quarterly GDP using monthly indicators for the United States and Germany, respectively. More recently, Andreou et al. (2013) explored the application of financial data within MIDAS frameworks to improve GDP growth forecasts in the U.S. Collectively, these studies have shown that incorporating mixed-frequency data significantly enhances predictive performance.

One persistent challenge in economic research is data availability. Many research institutions lack access to high-frequency GDP data but may have monthly figures for related indicators, such as the Cost of Imports (CIM). When GDP data is only available quarterly, the question arises: how can we effectively analyze relationships between high- and low-frequency variables?

A straightforward approach is to compute the arithmetic mean of high-frequency observations within each low-frequency period. However, this method assumes uniform contributions from all data points, which may not always be valid—more recent data often contain more relevant information. In such cases, assigning greater weight to recent observations is preferable. A naive linear regression treating each daily predictor value as a separate regressor would introduce excessive parameters, leading to high estimation uncertainty.

A compelling application of MIDAS is its use in modeling the risk-return trade-off, as demonstrated in the following regression model:

$$R_{t+1} = \mu + \gamma\hat{\sigma}_t^2 + \epsilon_{t+1} \tag{1}$$

where R_{t+1} is the excess return on the market in month $t+1$, and $\hat{\sigma}_t^2$ is theforecasted variance of returns for the same month $t+1$, based on informationknow at time t.

In most cases, additional observations of the high-frequency variable(s) become available after the most recent sample of the low-frequency dependent variable has been recorded. These extra data points can be incorporated into the analysis, enabling what is commonly referred to as "now-casting" in forecasting literature. The key advantage of this approach is that leveraging the most recent high-frequency information can enhance the model's predictive accuracy.

Unlike autoregressive (AR) models, which assume that past values are sampled at uniform intervals, MIDAS regressions accommodate regressors with varying sampling frequencies. As a result, they do not fall under the category of traditional AR models. Instead, MIDAS shares similarities with distributed lag models, yet it introduces distinctive features that set it apart. The distributed lag model follows a regression structure of the form:

$$Y_t = \beta_0 + B(L)X_t + \epsilon_t \tag{2}$$

where $B(L)$ is a finite or infinite lag polynomial operator, usually parameterized by a small set of hyperparameters.

This paper aims to study and implement the MIDAS regression model for GDP prediction using data from the Bureau of Economic Statistics for the year 2009. The structure of the paper is as follows:

- Section 2 presents a detailed overview of the methodology, including a comprehensive description of the MIDAS regression model, its underlying probability distributions, and the key parameters involved.

- Section 3 describes the dataset used in this study. We discuss the necessity of seasonal adjustments and the methods available for achieving them. Additionally, we highlight the importance of stationarizing a time series, addressing potential unit roots, and outlining techniques for their detection.

- Section 4 provides the R code used in our analysis, along with the corresponding results.

2 Methodology

A simple economic forecasting model involves using Ordinary Least Squares (OLS) regression to predict GDP based on its past values (lags) and another related economic variable, such as consumer prices, unemployment, or stock prices, along with the lag of that variable. The model can be expressed as:

$$Y_t = \beta_0 + \sum_{i=1}^{p} (\beta_i Y_{t-i} + \gamma_i X_{t-i}) + \epsilon_t \tag{3}$$

where Y_t represents GDP, X_t is the related economic variable, and ϵ_t is the error term with zero mean and constant variance at time t. Data up to time t can be used to forecast Y for time $t+1$.

However, time series data is often available at different frequencies. For example, GDP (Y) is typically published quarterly, whereas the related variable X might be available at a higher frequency. Data such as the Consumer Price Index (CPI) and unemployment figures are available monthly, and stock prices are recorded intraday. The most straightforward approach to handling different frequencies is to aggregate or average the higher-frequency data to match the lowest frequency. In this case, quarterly averages of the high-frequency variable would be used as the regressor X in the model. This method is akin to a restricted Least Squares regression, where the coefficients on high-frequency lags of X are assumed to be equal within the same quarter (the low-frequency period). However, this approach overlooks valuable information provided by the higher-frequency data.

An alternative approach is to regress Y on its own low-frequency lags, along with all high-frequency lags of X over the same time horizon:

$$Y_{t^{\mathrm{LF}}} = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t^{\mathrm{LF}}-i} + \sum_{k=1}^{mn} \gamma_i X_{t^{\mathrm{LF}}-k} + \epsilon_{t^{\mathrm{LF}}} \tag{4}$$

Here, m represents the number of high-frequency periods in each low-frequency period, and n is the number of low-frequency periods for which we include lags of X. Low and high-frequency time periods are denoted as t_{LF} and t_{HF}, respectively. For simplicity, we will omit the constant term and lag-dependent-variable terms in the subsequent discussion, as they remain unchanged across model specifications.

While this model allows for unique coefficient estimates for each high- and low-frequency observation, complications arise due to the large number of regressors. For example, if there are 60 daily stock price observations per quarter, a GDP forecasting model with only two quarter lags would require estimating 123 parameters. This can lead to low degrees of freedom, especially for time series with limited historical data, making statistical inference imprecise. Additionally, if high-frequency lags are strongly correlated, multicollinearity problems may occur. To address this, we could restrict the within-quarter effects to be the same across all quarters:

$$Y_{t^{\mathrm{LF}}} = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t^{\mathrm{LF}}-i} + \sum_{k=1}^{mn} \gamma_i X_{t^{\mathrm{LF}}-k} + \epsilon_{t^{\mathrm{LF}}} \tag{5}$$

However, even this simpler model would still require estimating 65 parameters for just two quarter lags in the stock price example.

2.1 MIDAS regressions

The Mixed-Data Sampling (MIDAS) regression model, introduced by Ghysels, Santa-Clara, and Valkanov (2004), is designed to preserve information from high-frequency data while mitigating issues such as parameter proliferation and multicollinearity. The MIDAS model is based on a distributed lag framework, but instead of estimating coefficients for each lag of the high-frequency regressor, it assigns weights to the lags using a polynomial function. The basic form of the MIDAS model is:

$$Y_{t^{\mathrm{LF}}} = \cdots + \gamma \sum_{k=1}^{mn} f(k;\theta) X_{t^{\mathrm{LF}}-k} + \epsilon_{t^{\mathrm{LF}}} \tag{6}$$

In this equation, $f(k;\theta)$ represents a polynomial function of the lag number k, and θ is a small set of hyperparameters that control the shape of the function. By weighting the lags according to this function, the model restricts the coefficient estimates on the weighted lags to a single value γ. As a result, the MIDAS model requires far fewer parameters than the models in Equations (4) or (5), needing only the parameters γ and θ.

The polynomial function $f(k;\theta)$ can take various forms, with two common choices being the beta formulation and the exponential Almon function. These functions are shown in graphs from Armesto, Engemann, and Owyang (2010), based on selected values for the hyperparameters θ_1 and θ_2, as illustrated in Figure 1.

Typically, the weights assigned to the lags decline as the high-frequency lag increases, often with a hump-shaped pattern in more recent periods. The overall downward slope reflects the intuition that the influence of past regressor data diminishes over time. However, recent movements in the regressor may require some time to impact the dependent variable, which accounts for the potential hump-shape in the weighting function.

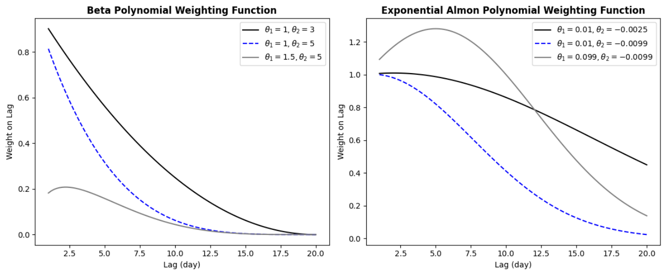

Figure 1:

A notable feature of MIDAS regressions is their ability to forecast the dependent variable in the current low-frequency period by incorporating high-frequency regressor data, a practice known as nowcasting. To achieve this, an additional term is added to the basic MIDAS model to account for high-frequency lags of the regressor between the present and the start of the current low-frequency period. The regression model for the d-th high-frequency period within the current low-frequency period is as follows:

$$Y_{t^{LF}|d} = \cdots + \gamma_1 \sum_{j=m-d+1}^{m} f(k;\theta_1) X_{t^{HF}+1-j} + \gamma_2 \sum_{k=d+1}^{n*m} f(k;\theta_2) X_{t^{HF}-k} + \epsilon_{t^{LF}} \tag{7}$$

This added term in Equation (7) ensures that the effects of data from the current low-frequency period are estimated separately from previous periods, using the parameters γ_1 and θ_1. Equations (6) and (7) are equivalent when $\gamma_1 = \gamma_2$ and $\theta_1 = \theta_2$.

3 Data

In this study, we apply MIDAS modeling techniques to economic data, specifically focusing on forecasting a country's Gross Domestic Product (GDP)—a quarterly time series variable—using monthly variables such as consumer prices and unemployment rates. We perform this analysis using data for the United States from 1980Q1 to 2017Q4.

Our data on real GDP, expressed in chained 2009 dollars, is sourced from the Bureau of Economic Statistics at a quarterly frequency. In contrast, all explanatory variables are available at a monthly frequency. From the Bureau of Labor Statistics, we use the urban Consumer Price Index (CPI), excluding food and energy, as a measure of core consumer prices, along with the civilian unemployment rate.

3.1 Seasonal Adjustment

Time series data often exhibit seasonal patterns that repeat annually but are not relevant for policy analysis. These patterns may reflect the seasonal weather cycle or annual holidays. For example, GDP tends to dip in the fourth quarter due to reduced production during the Christmas holidays, while unemployment typically decreases due to seasonal hiring. If we model these time series without adjusting for seasonality, the model could mistakenly indicate a positive relationship between GDP and unemployment, driven by these seasonal effects, which are not economically meaningful. This situation exemplifies omitted variable bias, leading to spurious correlations.

To address this, it is helpful to seasonally adjust the data to remove the effects of seasonal patterns. Two common methods for seasonal adjustment are X-12-ARIMA and X-13-ARIMA, developed by the US Census Bureau. These methods use moving averages to decompose time series into seasonal, trend, non-seasonal cycle, and idiosyncratic components, extracting the seasonal component from the data. In our study, most of the data comes from US statistical agencies and is already seasonally adjusted at the source, with the exception of consumer price data, which we seasonally adjust ourselves using X-13-ARIMA.

3.2 Stationarity and First Differences

When performing time series analysis, it is crucial to ensure that all variables are stationary, meaning they have a constant mean and variance over time. A non-stationary time series may exhibit a deterministic or natural time trend or may contain a unit root. A unit root is a parameter in the variable's underlying autoregressive (AR) process that influences its movement based on past values. If a variable has a unit root, any deviation from its equilibrium "steady state" value becomes permanent, meaning there is no time-invariant mean. Conversely, if the root is between zero and one, the variable will revert back to the equilibrium value over time, ensuring a constant mean. Non-stationary series with a root greater than one—an explosive root—are also non-stationary, though this phenomenon is rare in economic data.

Non-stationary time series can lead to spurious correlations. Seemingly significant relationships might actually be explained by a shared time trend or by similar AR processes that cause the variables to move together. For instance, a regression of infant mortality rates on the population of endangered whales could produce positive coefficients, but this relationship would be purely coincidental, driven by the fact that both variables are non-stationary and declining. This highlights the need to address non-stationarity before drawing any conclusions about relationships between variables.

There are common statistical tests to detect the presence of a unit root. These tests either treat the unit root as the null hypothesis, such as the Dickey-Fuller test, or as the alternative hypothesis, such as the KPSS test. Non-stationary variables, whether or not they have unit roots, can typically be made stationary by taking the first difference of the data or the first difference of the natural log if there is an exponential time trend. In this study, we make the data stationary by taking the first difference of the natural log for GDP and core CPI, as well as the first difference for the unemployment rate and stock prices.

4 Nowcasting real GDP

4.1 Method

Many time-sensitive policy decisions rely on data that is available infrequently. Often, policymakers receive this crucial data with a lag, and it may also be subject to revisions. Furthermore, different data sets are released at different times and at varying frequencies. For example, inflation and unemployment data are released monthly, while GDP data is released quarterly. Generating nowcasts for key economic variables using more readily available data can be a valuable tool for informing policy decisions.

To demonstrate the application of the MIDAS model, we develop a nowcasting model to examine the current levels of real GDP. Real GDP is a critical input for many economic decisions, such as monetary policy, government budgeting, and business planning. However, real GDP is typically released on a quarterly basis in most countries, whereas other economic indicators—such as industrial production, unemployment, inflation, and stock prices—are released at a monthly or higher frequency.

To evaluate the predictive power of our model, we compare the forecasts generated by the MIDAS model with those from a regular OLS model using a simple weighted average of the high-frequency variables. The model is estimated iteratively using an out-of-sample, rolling window forecast. The process works as follows: first, the model is estimated starting from 1980Q1 through to the quarter preceding the nowcast quarter. The nowcast is then generated using the available high-frequency data from that quarter. The nowcast is compared to the actual value using the root mean squared error (RMSE).

4.2 Unemployment

The graph below compares the nowcasts for real GDP generated using the unemployment rate with forecasts from both the MIDAS and OLS models. The actual real GDP values are shown in red, the MIDAS forecasts in green, and the OLS forecasts in blue. In this example, the MIDAS and OLS forecasts appear visually similar. However, comparing the RMSEs shows that the MIDAS model provides a slight improvement of about 5% over the OLS model.

4.3 Inflation

When using the MIDAS model with seasonally adjusted inflation as the explanatory variable, we find that it provides a significantly better nowcast for GDP compared to the OLS model. The chart below shows that while the OLS and MIDAS models are broadly similar in most periods, the MIDAS model tends to better predict declines in real GDP, particularly during recessions, such as those in 2001 and 2008. This is reflected in the lower RMSE for the MIDAS model, which demonstrates a 13% improvement over the OLS model.

5 Conclusion

MIDAS models are a valuable tool for forecasting low-frequency variables using high-frequency data. Unlike models that simply aggregate data to the lowest frequency, MIDAS models incorporate additional information from higher-frequency data while avoiding issues like parameter proliferation, multicollinearity, and low degrees of freedom that arise when estimating many high-frequency lags. This is achieved by assigning weights to the lags using a polynomial function governed by a small number of parameters, typically with decreasing weights as the lags move further into the past. Additionally, MIDAS models are effective for nowcasting the dependent variable in the current period.

In our application to economic data, we use monthly consumer prices and unemployment rates to forecast quarterly GDP. We ensure the variables are seasonally adjusted and stationary to avoid spurious correlations. By comparing MIDAS model forecasts to those from a simple OLS regression—where high-frequency variables are averaged to the lower frequency—we find that the MIDAS model consistently outperforms OLS for both CPI and unemployment, as indicated by their respective root mean squared errors.

References

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

[2] Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, **31**(2), 240–251.

[3] Armesto, M. T., Engemann, K. M., Owyang, M. T., *et al.* (2010). Forecasting with mixed frequencies. *Federal Reserve Bank of St. Louis Review*, **92**(6), 521–536.

[4] Clements, M. P., & 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.

[5] Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). The MIDAS touch: Mixed data sampling regression models. *UCLA Working Paper.*

[6] Giles, D. (2017). Explaining the Almon distributed lag model. `http://davegiles.blogspot.ca/2017/01/explaining-almon-distributed-lag-model.html`.

[7] Gomez-Zamudio, L. M., & Ibarra, R. (2017). Are daily financial data useful for forecasting GDP? Evidence from Mexico. *Economía*, **17**(2), 173–203.

[8] Granger IV, C. W., Hyung, N., & Jeon, Y. (2001). Spurious regressions with stationary series. *Applied Economics*, **33**(7), 899–904.

[9] Marcellino, M., & 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.

[10] Monsell, B. C. (2012). X-13-ARIMA-SEATS - a basic seasonal adjustment glossary. `https://www.census.gov/srd/www/x13as/glossary.html`.

[11] Monsell, B. C., Aston, J. A., & Koopman, S. J. (2003). Toward X-13. *ASA Proceedings, Business and Economic Statistics Section.*