Forecasting with Dynamic Factor Model and Mixed-Frequency Data

Mingdi Chen

A thesis submitted in partial fulfillment of the requirements for the degree of
Master of Philosophy

Discipline of Business Analytics
Business school
The University of Sydney

Supervisor: Associate Professor Jie Yin
Co-supervisor: Associate Professor Anastasios Panagiotelis

January 6, 2026

Statement of Originality

This is to certify that the content of this thesis is my own work. This thesis has not been submitted for any other degree or purpose.

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis and all sources have been acknowledged.

Mingdi Chen **Signature:** _____

Date: _____

Authorship Attribution Statement

Some parts of Section 1 of this thesis have been used in my thesis for the Master of Economic Analysis degree Chen (2022), and have been properly paraphrased. I designed the study, analysed the data, and wrote the drafts of the manuscript.

This thesis contains material previously used in my thesis for the Master of Economic Analysis Chen (2022). This material comprises Section 4. I reused RBA statement text data from the previous thesis and updated it through the year 2022.

Mingdi Chen **Signature:** _____

Date: _____

As supervisor for the candidature upon which this thesis is based, I confirm that the authorship attribution statements above are correct.

Jie Yin & Anastasios Panagiotelis **Signature:** _____

Date: _____

Artificial Intelligence

I acknowledge that generative artificial intelligence (AI) tools were used during the preparation of this thesis. Specifically, AI assistance was used for:

- Grammar checking and proofreading of draft chapters,
- Improving the clarity and expression of written English,
- Providing assistance with R and LaTeX coding and formatting,
- Troubleshooting syntax and resolving errors in LaTeX.

All work has been reviewed and appropriately edited by me to ensure academic integrity and compliance with university standards.

Mingdi Chen **Signature:** _____

Date: _____

Australian Government support

This research was supported by an **Enhanced Business School Research Scholarship (EBSRS)**.

Abstract

Macroeconomic forecasting plays a crucial role in shaping monetary and fiscal policies, guiding decision-making in both public and private sectors. Traditional economic indicators, often released at lower frequencies, are insufficient for capturing the rapid shifts in economic conditions, particularly in post-pandemic periods. This research aims to enhance macroeconomic forecasting by integrating sentiment analysis derived from higher-frequency textual data, such as central bank communications, into traditional econometric models. The research evaluates MIDAS, Kalman Filter-based Dynamic Factor Models (DFM), and MF-VAR to determine the effectiveness of incorporating sentiment alongside financial market indicators, such as the S&P/ASX 200 index.

The findings indicate that sentiment alone does not consistently outperform financial indicators but adds predictive value when combined with asset prices. In addition, integrating both sentiment and financial variables improves GDP forecasting accuracy, particularly at longer horizons. These findings underscore the value of high-frequency textual analysis in macroeconomic forecasting and highlight the complementary role of sentiment and financial market data in improving macroeconomic forecasting.

Keywords: macroeconomic forecasting, sentiment analysis, dynamic factor model, mixed-frequency data, central bank communication

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Associate Professor Anastasios Panagiotelis, and Associate Professor Jie Yin, for their invaluable guidance throughout this research. Their insightful comments and suggestions on earlier drafts, as well as their support in identifying relevant literature, developing the model, collecting and preparing data, adjusting estimation techniques, and interpreting the results, have been instrumental in shaping this thesis.

Most importantly, I am deeply thankful for their patience, understanding, and encouragement during a very difficult time in my personal life. Their support helped me persevere and complete this work despite the challenges I faced.

I am also grateful to Professor James Morley, Dr. Luke Hartigan, Dr. Ye Lu, and Dr. Dakyung Seong from the School of Economics for their constructive feedback and valuable suggestions during my proposal defense.

To all of you, thank you.

Contents

List of Tables

CHAPTER 1

Introduction

1.1 Background and motivation

Macroeconomic forecasting is one of the major aims of economic and econometric analysis, and also a critical tool for policymakers and financial institutions, as it provides insights into future economic status that inform decision-making processes. Accurate forecasts enable policymakers to design appropriate monetary and fiscal policies, and financial markets to predict economic trends. Therefore, substantial academic efforts have been dedicated to establishing foundational principles and developing tools for more efficient forecasting. In past decades, the discussion mainly covers the relative accuracy of macroeconometric models versus time series alternatives, the evolution of forecasting accuracy over time, the rationality of macroeconomic predictions, and the criteria for selecting a forecasting service (Fildes and Stekler (2002)).

Some literature focus on developing models that first seek to explain the underlying mechanisms of the economy, with forecasting as a secondary outcome. To some extent, this approach is optimal, as a model that accurately explains the economy can, in theory, also produce good forecasts. However, due to the complexity of the real economy and models required to capture its full dynamics, these forecasts may not always be accurate within the sample data, and are even less likely to perform well out-of-sample. Some other research focuses on models that do not aim for a comprehensive structural model but instead offer a simplified, reduced-form statistical description. These models often deliver better forecasting performance; however, their simplified nature makes it challenging to tell the real economic story needed to support and interpret the forecasts. Many economists and policymakers view this as a significant drawback (Carriero et al. (2019)).

One of the major challenges in macroeconomic forecasting is that data are not all sampled at same frequency. Most forecasting models, are generally based on a single frequency. In Australia, traditional economic indicators, such as GDP, employment rates, and inflation, are typically released quarterly or monthly with a lag and often subject to irregular revisions. In contrast, some text data, which can provide immediate insights into economic conditions such as news articles, social media posts, and RBA statements, are generated in higher frequencies and in real time. The main reason of including these text data into a forecasting model is they can enhance macroeconomic forecasts by capturing timely information that traditional indicators might miss. However, leveraging text data effectively requires advanced natural language processing

(NLP) techniques and a robust understanding of how to extract relevant signals from vast amounts of unstructured information.

Despite the increasing availability of high-frequency textual data in NLP, a little research has been conducted on integrating these sources systematically into formal macroeconomic forecasting models. Most existing work either treats text analysis and macroeconomic forecasting as separate domains or applies sentiment analysis in an ad hoc manner without evaluating its incremental forecasting value over traditional indicators. This lack of integration represents a significant gap in the literature, particularly in the Australian context where macroeconomic data are sparse and often published with delays.

1.2 Research question and contribution

Monetary policy communication and financial market conditions both play central roles in shaping expectations about the macroeconomy. Textual sentiment extracted from the RBA's monthly statements captures the qualitative, forward-looking component of policy signalling, while high-frequency financial indicators—such as movements in the ASX—provide market-based assessments of current and anticipated economic conditions. Despite their conceptual complementarity, these two information sources have rarely been jointly incorporated into a unified mixed-frequency forecasting framework for Australia. This motivates the central question of this thesis: To what extent can accounting for RBA sentiment (captured through textual analysis of high-frequency RBA statements) improve macroeconomic forecasts when integrated into a dynamic factor model with mixed-frequency data? Specifically, the thesis examines whether sentiment-based predictors and financial indicators enhance the out-of-sample forecasting accuracy of key macroeconomic variables, such as GDP, beyond what is achieved by models that rely solely on traditional macroeconomic data.

There are three key contributions of this thesis. First, it develops a framework that integrates sentiment extracted from textual data and financial conditions into a mixed-frequency dynamic factor model (DFM) for macroeconomic forecasting—an approach that is still relatively novel in the literature. Second, it conducts a rigorous empirical comparison of forecasting performance between models with different types of data (e.g. textual data only, traditional high-frequency financial indicator only, and both), using Australian post-pandemic data as a test set. Third, it highlights the practical implications of incorporating central bank communication into forecasting exercises, particularly in data-constrained environments in Australia.

In my prior work Chen (2022), Latent Dirichlet Allocation (LDA) was used to analyse the monthly statement of the RBA and investigate the macroeconomic effect of the central bank's communication. However, the unavailability of most key macroeconomic variables at monthly frequency in Australia led to the econometric model in Chen (2022) being specified using quarterly data, causing a loss of sample size and precision in estimation and forecasting.The new contribution of this thesis is the construction and estimation of mixed-frequency models that allow the monthly sentiment data to be

directly integrated with lower-frequency macroeconomic indicators. This improves both the timeliness and the potential accuracy of forecasts. Furthermore, the comparison between models with only financial predictors and those that also incorporate text-based sentiment is entirely novel in the Australian context.

The rest of this thesis will be organized as follows. In the next section, I provide literature from different related area, and more specific motivation as well as contribution based on that. Section 3 introduces the econometric methodology and the details of the estimation approaches. Section 4 gives description of both economics and text data. Section 5 gives the results analysis.

CHAPTER 2

Literature review

2.1 Textual sentiment as a predictor of economic activity

For macroeconomic forecasting, sentiment analysis will be able to capture the sentiment of the general public towards the economic shocks or policy changes. The study of Barbaglia et al. (2023) suggest that measures of economic sentiment can track closely business cycle fluctuations which indicate that these measures of sentiment are relevant predictors for macroeconomic variables. The aim of my research is to evaluate the sentiment from the RBA's monthly statement and economic-related content in newspapers, then use these measurements as predictors for macroeconomic forecasting. Specifically, the content of sentiment will be extracted from the text, and a sentiment score will be assigned to each words according to the scoring model of Shapiro et al. (2022) which is shown to have better predictive accuracy than existing "off-the-shelf" models. While Kalamara et al. (2022) similarly extract time-series economic information from newspaper text to improve macroeconomic forecasting, their study uses UK newspaper articles (The Guardian, The Daily Mail, and The Daily Mirror) from 1990 to 2019. In contrast, this thesis focuses primarily on central bank communication, specifically, the RBA's monthly statements, as a key source of textual information. Central bank communication provides signals about macroeconomic conditions and monetary policy intentions, helping shape market expectations (Blinder (2009)).

2.2 Central bank communication as textual sentiment

In today's monetary policy frameworks, communication plays a crucial role and has become one of the central tools used by monetary authorities to shape market expectations. It helps convey information about the economy's current condition and the future direction of monetary policy. Central bank communication is broadly defined as the information released by the central bank about its current and future policy objectives, current economic prospects, and likely paths for the decisions of future monetary policy (Hansen and McMahon (2016)). Historically, during the 1930s, institutions like the Federal Reserve operated with a high level of secrecy and offered little transparency to the public (Bholat et al. (2018)). The Reserve Bank of Australia (RBA) followed

a similar approach; as noted by a former Deputy Governor, "It was an article of faith in central banking that secrecy about monetary policy decisions was the best policy" (Lowe (2020)). This stance has changed considerably over recent decades. Since the early 1990s—and more notably following the 2008 global financial crisis—central banks have increasingly recognized the value of open communication. They now regularly publish their policy goals and economic assessments to promote greater transparency and public understanding (Haldane and McMahon (2018)).

For computational feasibility and decreasing the noise ratio, Kalamara et al. (2022)'s study made restrictions on the news type, they used only regular news, editorials and commentaries. Therefore, for the same reason, the communication type in my research will be Governor's monthly statement, since these statements are released regularly and formed in fixed structure. Another reason of this choice is that the RBA statement is extremely well-structured especially after Lowe became the governor. The first few paragraphs are mainly about **Global economy**, the middle part is **Domestic economy**, and the last few paragraphs are about **Forward guidance**. Thus, three sentiment indicators will be extracted from three part separately and served as high-frequency textual data in models.

In Australia context, He (2021) examines how RBA's announcement impacts the Australian equity market, and found that the information effect could be present in other forms of RBA's communication. However, research integrating RBA communication into structured forecasting models remains limited. This thesis addresses this gap by extracting structured sentiment from the RBA Governor's monthly statements and assessing their predictive power for core macroeconomic indicators.

Recent years have seen rapid growth in research applying textual analysis to central bank communication, particularly in the aftermath of the COVID-19 pandemic, when uncertainty and policy interventions increased dramatically. A central contribution in this area comes from Ter Ellen et al. (2022), whose work demonstrates how narrative and sentiment extracted from monetary policy communication shape financial markets and macroeconomic expectations. They combined textual features from media coverage with high-frequency market data to identify how communication channels amplify or dampen policy signals. Their findings suggest that textual narratives can explain variation in financial market responses beyond what is captured by conventional numerical policy surprises. Their research highlights the value of structured textual indicators as complementary tools for understanding and forecasting macroeconomic dynamics, which is also the motivation underlying this thesis.

Some other post-pandemic studies further underscore the role of communication during periods of higher uncertainty. Several studies (e.g. Benchimol et al. (2025)) show that sentiment extracted from central bank speeches, press conferences, or meeting minutes can predict inflation expectations or real economic activity more effectively when traditional indicators become less informative. These findings collectively support the view that textual indicators, particularly those capturing tone, uncertainty, and forward-looking narratives have become increasingly important components of modern empirical macroeconomics. This growing body of research provides strong support

for incorporating structured sentiment from RBA communication into macroeconomic forecasting frameworks.

2.3 Applications of NLP in macroeconomic text analysis

Bidirectional Encoder Representations from Transformers (BERT) is a language model introduced by Google in 2018, it will help to process and quantify the linguistic information to construct several variables which can be used as predictors in the econometric model. The input of BERT will be the text in sentence level from RBA statement. I'm proposing to put them into one same pool for pre-training as in Barbaglia et al. (2024). The reason is that RBA statements are well-structured and consistent in tone, each paragraph contains a certain content in almost all statements (e.g., the last paragraph is usually about forward guidance).

After the pre-training, the inputs of the model are the text, with these inputs, BERT will assign a sentiment score for each related sentence related to a certain topic range from -1 to 1, then obtain an average score of the indicator for each text. The output will be aggregate to the monthly frequency by averaging the values within the week. In my application, I will mainly focus on the sentiment related to the three topics listed above: global economy, domestic economy, and forward guidance. Thus, in total, I will have three sentiment indicators, and these sentiment indicators will then be used in my forecasting model as predictors.

Ellingsen et al. (2022) compared news-based indicators with the widely used FRED-MD data in predicting U.S. macroeconomic variables. They find that when processed using modern NLP techniques, news media data can provide forecasting gains that are both economically meaningful and complementary to standard macroeconomic indicators. Their results highlight two key insights relevant to this thesis: first, textual information contains forward-looking content that may not be present in quantitative macro data; and second, machine-readable narratives extracted from structured text can materially improve the accuracy of macroeconomic forecasts. This evidence further motivates the construction of sentiment indicators from RBA communication as an additional source of predictive information.

2.4 Modeling Mixed-Frequency Data in Macroeconomics

Macroeconomic forecasting frequently involves integrating information sampled at different frequencies—e.g., quarterly GDP and monthly or even daily financial indicators. Most traditional econometric models typically assume data is sampled at a uniform frequency, but this assumption can lead to loss of information or mismeasurement when dealing with real-world data sources. A simple approach based on this is up-sampled low-frequency variables to high frequency by interpolation (Kalamara et al. (2022)), this

method is also applied in my previous research Chen (2022). However, interpolation will cause the loss of precision when doing forecasting. To deal with, a range of econometric techniques have been developed to directly accommodate mixed-frequency data, the following three approaches are widely-used and will be estimated in this thesis to make comprasion: Mixed Data Sampling (MIDAS), Mixed-Frequency VAR (MF-VAR), and Dynamic Factor Models (DFM) estimated via state-space methods.

2.4.1 Mixed Data Sampling (MIDAS)

The MIDAS approach, originally introduced by Ghysels et al. (2004), is a regression-based framework that allows low-frequency dependent variables (e.g., quarterly GDP) to be forecast using high-frequency regressors (e.g., monthly sentiment or financial indicators) without needing to aggregate or interpolate the high-frequency data. MIDAS models achieve this by applying 'restricted' lag polynomials, such as the Almon lag, to control the influence of high-frequency lags while maintaining parsimony (Kuzin et al. (2011)).

The MIDAS framework has gained wide acceptance for its computational simplicity and strong empirical performance in nowcasting and short-term forecasting (Andreou et al., 2010; Foroni et al., 2015). Its application has been extended beyond macro indicators to incorporate high-frequency financial data and even unstructured information like sentiment scores. Barbaglia et al. (2023) utilize a MIDAS model with sentiment extracted from news articles to forecast UK GDP and find that it outperforms traditional autoregressive benchmarks.

However, MIDAS models face certain limitations. The use of fixed-weight lag structures may underfit in settings where dynamics are more complex or where relationships are non-linear. Moreover, MIDAS is primarily designed for univariate forecasting problems and does not easily accommodate endogenous interactions among predictors—a gap filled by more flexible frameworks like MF-VAR Kuzin et al. (2011).

2.4.2 Mixed-Frequency VAR (MF-VAR)

Mixed-Frequency VAR (MF-VAR) models extend the conventional vector autoregression framework to accommodate variables observed at different sampling frequencies. Rather than upsampling or temporally interpolating high-frequency indicators, MF-VARs explicitly model the underlying latent monthly or quarterly dynamics that give rise to the observed mixed-frequency data. In this sense, MF-VARs can be viewed as a specific class of state-space models in which the latent state evolves at the highest frequency, while the measurement equation links the state to lower-frequency observations. This state-space representation was formalised in Schorfheide and Song (2015), who show how mixed-frequency macroeconomic data can be handled naturally within such a structure.

An important advantage of MF-VARs is that they allow high- and low-frequency variables to interact contemporaneously and dynamically, making it possible to model feed-

back effects and cross-series propagation mechanisms that are absent in MIDAS regressions. Marcellino and Schumacher (2010) find that MF-VARs often outperform both traditional VARs and MIDAS models at medium- and long-horizon forecasts, particularly in settings where macroeconomic shocks diffuse gradually across time and across frequencies.

However, MF-VARs come with costs as well. Their flexibility increases parameter dimensionality, especially when many high-frequency lags are included. Since estimation typically proceeds through Bayesian methods or EM algorithms, the curse of dimensionality can lead to overfitting in small samples and requires careful regularisation or shrinkage.

The next subsection introduces the general state-space framework and Dynamic Factor Models (DFMs), which encompass MF-VARs as a special case but provide additional structure to reduce dimensionality and extract common latent components in large systems.

2.4.3 Dynamic Factor Models and State-Space Approaches

An alternative way to deal with mixed-frequency data is State-Space model (SSM) which includes a two equation system: one state equation describes how state evolve over time and one measurement equation which connect observed data to unobserved state process. SSM can handle various complexities in time series data, such as missing data, and time-varying parameters (Bai et al. (2013)). A Dynamic Factor Model (DFM) can be seen as a specific type of state-space model that assumes that the observed data is driven by a few underlying unobserved factors that evolve over time. These factors capture the common dynamics across multiple macroeconomic series. The model is commonly expressed in state-space form and estimated via the Kalman filter, which can more efficiently handles missing data, irregular publication lags, and mixed-frequency observations than MIDAS and MF-VAR (Stock and Watson (2011)). Also, there is a trade-off between over-fitting and lack of information in traditional econometrics models, while factor models can mitigate this overfitting (Bernanke et al. (2005)).

The foundational contributions of Stock and Watson (2002) formalized the static and dynamic versions of the factor model, leading to widespread applications in macroeconomics. They use DFM to forecast a macroeconomic time series variable using a large number of predictors, their finding shows that DFM with a smaller number of factors perform well at most cases, while in a few cases the performance is not that good. Mariano and Murasawa (2010) extend DFMs to approximate a monthly measure of real GDP from various indicators. More recently and related, Okuneva et al. (2024) develop mixed-frequency DFMs that combine textual and traditional indicators, emphasizing their flexibility and superior real-time performance.

DFM has two main advantages: (1) parsimony in handling large datasets, and (2) the ability to incorporate unobserved information through latent factors. However, their accuracy depends on the correct specification of factor dynamics and the assumption that all relevant information is reflected in the chosen observable series.

In a direct comparison, Kuzin et al. (2011) compare MIDAS with mixed-frequency vector-autoregressive (MF-VAR), they indicate that both model shows specific advantages and disadvantages. For MIDAS, the non-linear distributed lag function may be overly restrictive, while MF-VAR has no restriction on dynamics so will suffer from the curse of dimensionality. Their empirical findings show similarly that MIDAS performs better for short horizon while MF-VAR for long horizon. Bai et al. (2013) finds that DFMs estimated via Kalman filtering often outperform MIDAS models in high-dimensional settings where latent common factors explain most of the variation. In contrast, MIDAS is shown to be more efficient in small samples and short-term forecasting due to its simplicity. Consequently, the relative performance of each model depends on sample size, forecast horizon, and the richness of the data environment. Thus, DFM estimated via Kalman filter will serve as the main forecasting model in my research, while MIDAS and MF-VAR will also be presented for comparison.

CHAPTER 3

Methodology

The main econometric methods used in this project is a Dynamic factor model (DFM) in State-Space form introduced in Bai et al. (2013), which will be estimated by Kalman filter. At the same time, two other mixed-frequency models like MIDAS and mixed-frequency VAR will also be adopted as benchmarks. This section will first give introduction of the State-space model (SSM) and Kalman filter, then extend it to the special cases of having mixed-frequency data.

3.1 Mixed-frequency models

3.1.1 State-space model

State-space models provide a flexible and coherent framework for analysing dynamic systems in which observable variables depend on unobserved (latent) states that evolve over time. A standard linear Gaussian state-space system consists of two equations: the state equation, which governs the evolution of the latent state, and the observation equation, which links the latent state to the observed data.

Let

- $\mathbf{z}_t \in \mathbb{R}^{n_z}$ denote the vector of observed variables at time t,
- $\boldsymbol{\alpha}_t \in \mathbb{R}^{n_\alpha}$ denote the latent state vector,
- $\mathbf{F} \in \mathbb{R}^{n_\alpha \times n_\alpha}$ denote the state transition matrix,
- $\mathbf{H} \in \mathbb{R}^{n_z \times n_\alpha}$ denote the observation (or loading) matrix.

The system is given by:

State equation

$$\boldsymbol{\alpha}_{t+1} = \mathbf{F}\boldsymbol{\alpha}_t + \boldsymbol{\zeta}_{t+1}, \tag{3.1}$$

where

$$\boldsymbol{\zeta}_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}),$$

and $\mathbf{Q} \in \mathbb{R}^{n_\alpha \times n_\alpha}$ is the covariance matrix of state disturbances. The state equation describes the law of motion of the latent system, such as factor evolution or high-frequency VAR dynamics.

Observation equation

$$\mathbf{z}_t = \mathbf{H}\boldsymbol{\alpha}_t + \boldsymbol{\omega}_t, \tag{3.2}$$

where

$$\boldsymbol{\omega}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}),$$

and $\mathbf{R} \in \mathbb{R}^{n_z \times n_z}$ is the covariance matrix of measurement errors.

Shock assumptions

We assume:

- $\boldsymbol{\zeta}_t$ and $\boldsymbol{\omega}_t$ are serially uncorrelated,
- $\mathbb{E}[\boldsymbol{\zeta}_t \boldsymbol{\omega}_s'] = \mathbf{0}$ for all t, s, i.e., the two shocks are mutually independent at all lags,
- both shock vectors follow multivariate Normal distributions.

This structured representation allows the use of the Kalman filter for estimation, smoothing, and forecasting, and serves as a unifying framework for many macroeconomic time-series models—including mixed-frequency VARs and Dynamic Factor Models (DFMs)—because latent state vectors can be designed to encode the relevant dynamics at the desired frequency. In addition, the expanding window method is used in the PCA to avoid data leakage, which will be introduced in detail later.

3.1.2 General Dynamic factor model

Start from a general dynamic factor model, let $\mathbf{f}$ be a $n_f \times 1$ dimensional vector process satisfying

$$\mathbf{f}_{t+1} = \sum_{l=1}^{p} \Phi_l \mathbf{f}_t + \eta_{t+1} \quad \forall t = 1, \ldots, T, \tag{3.3}$$

where Φ_l are $n_f \times n_f$ matrices, the eigenvalues of the companion matrix lie strictly inside the unit circle, and η is an i.i.d. zero mean Gaussian error process with diagonal covariance matrix $\Sigma_\eta = \text{diag}(\sigma^2_{\eta i}, i = 1, \ldots, n_f)$. n_f is the number of factors. Observed data will relate to the factors through factor loading:

$$\mathbf{w}_{it} = \boldsymbol{\gamma_i}' \mathbf{f}_t + u_{it} \tag{3.4}$$

3.1.3 Mixed-frequency Dynamic factor model

When extended to mixed-frequency case, there are two types of data, low frequency (observed at integer $1, 2, \ldots, T$) and high frequency (observed at $1, 1 + (1/m), 1 + (2/m), \ldots, (T-1) + ((m-1)/m), T$), where m is the number of high frequency periods in one low frequency period (e.g. m=3 for monthly/quarterly data). Note that here I just assume there is only one low-frequency process y, with multiple high-frequency process $\mathbf{x}_t = (x_{1t} x_{2t} \cdots x_{nt})$.

$$\mathbf{f}_{t+j/m} = \sum_{l=1}^{p} \Phi_l \mathbf{f}_{t+(j-l)/m} + \eta_{t+j/m} \quad \forall t = 1, \ldots, T, \quad j = 0, \ldots, m-1, \tag{3.5}$$

If the LF process were observed at HF, it would relate to the factors as follows:

$$y^*_{t+j/m} = \boldsymbol{\gamma}_1' \mathbf{f}_{t+j/m} + u_{1,t+j/m} \quad \forall t, \quad j = 0, \ldots, m-1, \tag{3.6}$$

where y^* denotes the process which is not directly observed and $\boldsymbol{\gamma}_1$ is a $n_f \times 1$ vector of factor loadings. The error process $u_{1,t+j/m}$ has an AR(k) representation:

$$d_1(L^{1/m}) u_{t+j/m} = \varepsilon_{t+j/m}, \quad d_1(L^{1/m}) = 1 - d_{11}L^{1/m} - \ldots - d_{k1}L^{k/m}, \tag{3.7}$$

where the lag operator $L^{1/m}$ applies to HF data, that is $L^{1/m} u_t = u_{t-1/m}$.

The observed LF process y relates to the process y^* via a linear aggregation scheme if y is a flow variable:

$$y^c_{t+j/m} = y^c_{t+(j-1)/m} + 1/m y^*_{t+j/m}, \tag{3.8}$$

where y_t is equal to y^c_t for integer t, and is not observed otherwise, y^c_t is a cumulator variable. If y is a stock variable, y^c_t will only take the value equal to y_t when it's observed, and others ignored.
The HF process $\mathbf{x}_{t+j/m}$ relates to the factors as follows:

$$\mathbf{x}_{t+j/m} = \boldsymbol{\gamma}_2' \mathbf{f}_{t+j/m} + \boldsymbol{u}_{t+j/m} \quad \forall t, \quad j = 0, \ldots, m-1, \tag{3.9}$$

where $\boldsymbol{\gamma}_2$ is a $n_f \times (n-1)$ matrix and:

$$d_2\left(L^{1/m}\right) u_{2,t+j/m} = \varepsilon_{2,t+j/m}, \quad d_2\left(L^{1/m}\right) = 1 - d_{12}L^{1/m} - \ldots - d_{k2}L^{k/m}. \tag{3.10}$$

As usual in latent factor models, factor loadings γ_1, γ_2 and the parameters of the factor dynamics are subject to identification restrictions.

3.1.4 The State-Space representation of a dynamic factor model

State equation

$$\boldsymbol{\alpha}_{t+1} = \mathbf{F}\boldsymbol{\alpha}_t + \boldsymbol{\zeta}_{t+1}, \tag{3.11}$$

where the state vector is

$$\boldsymbol{\alpha}_t = \left(\mathbf{f}_t', \mathbf{f}_{t-1}', \ldots, \mathbf{f}_{t-p+1}', \mathbf{u}_t', \mathbf{u}_{t-1}', \ldots, \mathbf{u}_{t-k+1}'\right)',$$

with:
$\mathbf{f}_t \in \mathbb{R}$: vector of common factors, $\mathbf{u}_t = (u_{1t}, \ldots, u_{nt})' \in \mathbb{R}^n$: idiosyncratic shocks, p: factor lag order, k: idiosyncratic lag order.
Thus the dimension of the state vector is

$$\dim(\boldsymbol{\alpha}_t) = p + nk.$$

Companion matrix for the factor VAR(p):

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{p \times p}.$$

Companion matrix for the idiosyncratic AR(k) processes:

$$\mathbf{\Psi_i} = \begin{bmatrix} \psi_{i1} & \psi_{i2} & \cdots & \psi_{ik-1} & \psi_{ik} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{k \times k}.$$

So the **Transition Matrix** is:

$$\mathbf{F} = \text{diag}\left(\mathbf{\Phi},\ \mathbf{\Psi_1},\ \ldots,\ \mathbf{\Psi_k}\right) \in \mathbb{R}^{(p+nk)\times(p+nk)}.$$

and ζ_{t+1} is $(p + n * k) \times 1$ Gaussian White noise idiosyncratic error with variance-covariance matrix $\mathbf{Q}$.

Observation equation

$$\mathbf{z}_t = \mathbf{H}\alpha_t \tag{3.12}$$

where $\mathbf{z}_t$ denotes an $n \times 1$ vector of variables observed at time t, and observed data vector in the mixed-frequency context is

$$\mathbf{z}_t = (y_t, x_{2t}, \ldots, x_{nt})'$$

Here, y_t denotes the low-frequency variable of interest and the remaining $(n-1)$ elements $x_{2t}, \ldots, x_{nt}$ denote the high-frequency indicators, so that $\mathbf{z}_t \in \mathbb{R}^n$.

$\mathbf{H}$ is $n \times (p + n * k)$ observation matrix, where

$$\mathbf{H} = \begin{bmatrix} \gamma_1 & \mathbf{O}_{p-1} & 1 & \mathbf{O}_{p-1} & 0 & \cdots & 0 & \mathbf{O}_{p-1} \\ \gamma_2 & \mathbf{O}_{p-1} & 0 & \mathbf{O}_{p-1} & 1 & \cdots & 0 & \mathbf{O}_{p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_n & \mathbf{O}_{p-1} & 0 & \mathbf{O}_{p-1} & 0 & \cdots & 1 & \mathbf{O}_{p-1} \end{bmatrix}$$

zero blocks of dimension $n \times (p-1)$ are inserted right to each column of the identity matrix $\mathbf{I}_n$.
Here we just assume there is no exogenous variable and no measurement error. Once in state-space form, the Kalman filter can be applied to efficiently estimate the unobserved factors in real-time, even as new data becomes available. This makes it possible to update predictions dynamically and handle missing data effectively.

3.2 Kalman filter

The Kalman filter is a powerful statistical tool widely used in econometrics for estimating the dynamic behaviour of time series data. It has found extensive applications in economics due to its ability to provide efficient, real-time estimates of unobserved variables, such as the underlying state of the economy or other latent factors, by combining noisy observational data with a model of the system's dynamics. Particularly, Kalman filter is useful for dealing with models that evolve over time, such as state-space models, where the relationships between variables are subject to change. It allows to

update estimates as new data becomes available, making it a crucial tool for real-time forecasting.

The Kalman filter operates through a series of steps that can be divided into two main stages: the Prediction (Time Update) stage and the Update (Measurement Update) stage.

1. **Initialization**

   - Initial state estimate α_0
   - Initial covariance estimate $\mathbf{P}_0$

2. **Prediction stage**

   The prediction stage makes predictions of the state components based on information up to time $t-1$.

   - State prediction
   $$\alpha_{t|t-1} = \mathbf{F}\alpha_{t-1|t-1} \tag{3.13}$$
   where $\alpha_{t|t-1}$ is the predicted state vector at time t, $\mathbf{F}$ is the transition matrix.

   - Covariance prediction
   $$\mathbf{P}_{t|t-1} = \mathbf{F}\mathbf{P}_{t-1|t-1}\mathbf{F}' + \mathbf{Q} \tag{3.14}$$
   where $\mathbf{P}_{t|t-1}$ covariance matrix of the state estimate at time t, $\mathbf{Q}$ is the covariance matrix of ζ_t.

3. **Updating stage**

   The updating stage makes predictions based on all information up to time t.

   - State update
   $$\alpha_{t|t} = \alpha_{t|t-1} + \mathbf{K}_t(z_t - \mathbf{H}\alpha_{t|t-1}) \tag{3.15}$$
   where $\mathbf{K}_t$ is the kalman gain defined as
   $$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}'(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' + \mathbf{R})^{-1} \tag{3.16}$$
   where $\mathbf{R}$ is the measurement noise covariance which is set to zero here.

   - Covariance update
   $$\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_{t|t-1}\mathbf{H})\mathbf{P}_{t|t-1} \tag{3.17}$$

4. **Recursion**

   Repeat the Prediction and Update stages for each subsequent time step as new measurements become available. Kalman filter can deal with missing values. If $\mathbf{z}_t$ is missing for any observation, then the Kalman filter skips the update step, which is equivalent to inflating the measurement variance to infinity, causing the Kalman gain to be zero.

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} \tag{3.18}$$
$$\mathbf{K}_t = 0 \tag{3.19}$$
$$\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}' \propto \infty \tag{3.20}$$

To estimate a model with the Kalman filter we use maximum likelihood, the log-likelihood is given by:

$$\mathcal{L} = -\frac{1}{2}\sum_{t=1}^{T}\left(\mathbf{z}_t - \mathbf{H}\alpha_{t|t-1}\right)'\left(\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}'\right)^{-1}\left(\mathbf{z}_t - \mathbf{H}\alpha_{t|t-1}\right) - \frac{1}{2}\sum_{t=1}^{T}\ln\left|\mathbf{H}\mathbf{P}_{t|t-1}\mathbf{H}'\right| \tag{3.21}$$

3.2.1 Mixed-frequency case: Periodic Kalman Filter

To deal with mixed-frequency issue, here extend to periodic Kalman Filter. Here is the derivation of steady state Kalman filter formula for this periodic state space model according to Assimakis and Adam (2009) and Bai et al. (2013).
Observation equation:

$$\mathbf{z}_t^j = \mathbf{H}_j\alpha_{t+j/m}, \begin{cases} \mathbf{z}_t^j = (y_t, x_{2t}, x_{3t}, \ldots, x_{nt})' & j = m \\ \mathbf{z}_t^j = (x_{2t+j/m}, x_{3t+j/m}, \ldots, x_{nt+j/m})' & 1 \le j \le m-1 \end{cases} \tag{3.22}$$

the state vector

$$\alpha_{t+j/m} = \left(\mathbf{f}_{t+j/m}, \ldots, \mathbf{f}_{t+(j-p+1)/m}, \mathbf{U}_{t+j/m}, \ldots, \mathbf{U}_{t+(j-k+1)/m}\right)',$$

where $\mathbf{U}_{t+j/m} = \left(u_{1,t+j/m}, \ldots, u_{n,t+j/m}\right)'$, observation matrix $\mathbf{H}$ is $(n-1)\times(p+n*k)$ for $1 \le j \le m-1$ with the first line dropped

$$\mathbf{H}_m = \begin{bmatrix} \gamma_2 & \mathbf{O}_{p-1} & 0 & \mathbf{O}_{p-1} & 1 & \cdots & 0 & \mathbf{O}_{p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_n & \mathbf{O}_{p-1} & 0 & \mathbf{O}_{p-1} & 0 & \cdots & 1 & \mathbf{O}_{p-1} \end{bmatrix}$$

and same as before for $j = m$

$$\mathbf{H}_j = \begin{bmatrix} \gamma_1 & \mathbf{O}_{p-1} & 1 & \mathbf{O}_{p-1} & 0 & \cdots & 0 & \mathbf{O}_{p-1} \\ \gamma_2 & \mathbf{O}_{p-1} & 0 & \mathbf{O}_{p-1} & 1 & \cdots & 0 & \mathbf{O}_{p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_n & \mathbf{O}_{p-1} & 0 & \mathbf{O}_{p-1} & 0 & \cdots & 1 & \mathbf{O}_{p-1} \end{bmatrix}$$

The transition equation is

$$\alpha_{t+j/m} = \mathbf{F}\alpha_{t+(j-1)/m} + \zeta_{t+j/m} \tag{3.23}$$

$\mathbf{F}$ is $(p + n * k) \times (p + n * k)$ transition matrix same as before, where

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi_1 & \Phi_2 & \cdots & \Phi_{p-1} & \Phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{p \times p}.$$

$$\boldsymbol{\Psi_i} = \begin{bmatrix} \psi_{i1} & \psi_{i2} & \cdots & \psi_{ik-1} & \psi_{ik} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \in \mathbb{R}^{k \times k}.$$

$$\mathbf{F} = \text{diag}\left(\boldsymbol{\Phi},\ \boldsymbol{\Psi_1},\ \ldots,\ \boldsymbol{\Psi_k}\right) \in \mathbb{R}^{(p+nk)\times(p+nk)}.$$

$\zeta_{t+j/m} = (\eta_{t+j/m}, \varepsilon_{1,t+j/m}, \ldots, \varepsilon_{n,t+j/m})'$. Let $\mathbf{Q}$ denote the variance-covariance matrix of $\zeta_{t+j/m}$.
This state space model is periodic since it cycles to the data release pattern that repeats every m periods. We denote the steady state covariance matrix of $\alpha_{t+j/m|t+(j-1)/m}$ as $\mathbf{P}_{j|j-1}$, we have forecasting equation

$$\alpha_{j|j-1} = \mathbf{F}\alpha_{j-1|j-1} \tag{3.24}$$

$$\mathbf{P}_{j|j-1} = \mathbf{F}\mathbf{P}_{j-1|j-1}\mathbf{F}' + \mathbf{Q} \tag{3.25}$$

and the updating equation

$$\alpha_{j|j} = \alpha_{j|j-1} + \mathbf{P}_{j|j-1}\mathbf{H}_j'(\mathbf{H}_j\mathbf{P}_{j|j-1}\mathbf{H}_j')^{-1}(\mathbf{z}_j - \mathbf{H}_j\alpha_{j|j-1}) \tag{3.26}$$

where the steady-state covariance matrix of $\alpha_{t+j/m|t+(j-1)/m}$

$$\mathbf{P}_{j|j} = (\mathbf{I} - \mathbf{K}_{j|j-1}\mathbf{H}_j)\mathbf{P}_{j|j-1} = \mathbf{P}_{j|j-1} - \mathbf{P}_{j|j-1}\mathbf{H}_j'(\mathbf{H}_j\mathbf{P}_{j|j-1}\mathbf{H}_j')^{-1}\mathbf{H}_j\mathbf{P}_{j|j-1} \tag{3.27}$$

for $j = 1, \cdots, m-1$

$$\mathbf{P}_{1|m} = (\mathbf{I} - \mathbf{K}_{m|m-1}\mathbf{H}_m)\mathbf{P}_{m|m-1} = \mathbf{P}_{m|m-1} - \mathbf{P}_{m|m-1}\mathbf{H}_m'(\mathbf{H}_m\mathbf{P}_{m|m-1}\mathbf{H}_m')^{-1}\mathbf{H}_m\mathbf{P}_{m|m-1} \tag{3.28}$$

for $j = m$ which will satisfy $\alpha_{j|j-1} \equiv \alpha_{j+m|j-1+m}$

the steady-state Kalman gain is given by:

$$\mathbf{K}_{j|j-1} = \mathbf{P}_{j|j-1}\mathbf{H}_j'(\mathbf{H}_j\mathbf{P}_{j|j-1}\mathbf{H}_j')^{-1}$$

where $\mathbf{K}_{j|j-1} \equiv \mathbf{K}_{j+m|j-1+m}$

For state vector, we have

$$\hat{\alpha}_{t+j/m|t+j/m} = E[\alpha_{t+j/m}|\mathbf{z}_t^j, \mathbf{z}_t^{j-1}, \cdots, \mathbf{z}_{t-1}^m, \mathbf{z}_{t-2}^m, \cdots] \tag{3.29}$$

and the filtered states are

$$\hat{\alpha}_{t+j/m|t+j/m} = \mathbf{A}_{j|j-1}\hat{\alpha}_{t+(j-1)/m|t+(j-1)/m} + \mathbf{K}_{j|j-1}\mathbf{z}_t^j \tag{3.30}$$

where $\mathbf{A}_{j|j-1} = \mathbf{F} - \mathbf{K}_{j|j-1}\mathbf{H}_j\mathbf{F}$ represents an adjustment to the state transition matrix $\mathbf{F}$ that accounts for the new information gained from the observation at time j.

If we only interested in forecasting at LF by using all available LF and HF data, first according to Assimakis and Adam (2009) we have

$$\hat{\alpha}_{t+k|t+k} = [\tilde{\mathbf{A}}_1^m]^k\hat{\alpha}_{t|t} + \sum_{i=1}^{m}\sum_{j=1}^{k}[\tilde{\mathbf{A}}_1^m]^{k-j}\tilde{\mathbf{A}}_{i+1}^m\mathbf{K}_{i|i-1}\mathbf{z}_{t+j-1}^i \tag{3.31}$$

where k is integer and

$$\tilde{\mathbf{A}}_j^i = \begin{cases} \tilde{\mathbf{A}}_{i|i-1}\tilde{\mathbf{A}}_{i-1|i-2}\cdots\tilde{\mathbf{A}}_{j|j-1} & \text{for } i \geq j \\ \mathbf{I} & \text{for } i < j \end{cases}$$

According to Bai et al. (2013), we can iterate equation (32) backwards then have

$$\hat{\alpha}_{t|t} = \sum_{j=0}^{\infty}\sum_{i=1}^{m}[\tilde{\mathbf{A}}_1^m]^j\tilde{\mathbf{A}}_{i+1}^m\mathbf{K}_{i|i-1}\mathbf{z}_{t-j}^i = \sum_{j=0}^{\infty}[\tilde{\mathbf{A}}_1^m]^j\mathbf{K}_{m|m-1}\begin{pmatrix} y_{t-j} \\ x_{t-j} \end{pmatrix} + \sum_{j=0}^{\infty}\sum_{i=1}^{m-1}[\tilde{\mathbf{A}}_1^m]^j\tilde{\mathbf{A}}_{i+1}^m\mathbf{K}_{i|i-1}(x_{t-1-j+i/m}) \tag{3.32}$$

from where we can construct the forecast for LF process by using both LF and HF data as

$$E_t[y_{t+h}] = \mathbf{H}_{m,1}\mathbf{F}^{mh}\hat{\alpha}_{t|t}$$

where $\mathbf{H}_{m,1}$ denotes the first line of $\mathbf{H}_m$

3.3 Benchmark models

Three models are used as the empirical evaluation undertaken in this research. As mentioned earlier, MIDAS and MF-VAR show specific advantages and disadvantages, so it is reasonable to include both of them to make a comparison. AR(p) model can served as benchmark model since it can be considered as forecasting without high-frequency sentiment indicators.

3.3.1 MIDAS

A general (restricted) MIDAS with autoregressive dynamics used in Kuzin et al. (2011):

$$y_{t+h} = \beta_y\sum_{j=0}^{K_y}w_j(\theta_y)y_{t-j} + \beta_x\sum_{j=0}^{K_x}w_j(\theta_x^1)^j x(\theta_x^2)_{t-j} + \varepsilon_{t+h}, \tag{3.33}$$

where K_y and K_x are lag order, and $w_j(\theta_y)$, $w_j(\theta_x^1)$ follow an exponential Almon lag scheme where

$$w_j(\theta_y) = \frac{exp(\theta_{y1}k + \theta_{y2}k^2)}{\sum_{k=0}^{K_y}exp(\theta_{y1}k + \theta_{y2}k^2)}$$

and

$$x(\theta_x^2)_{t-j} = \sum_{k=0}^{m-1} w_k(\theta_x^2) L^{k/m} x_{t-k/m},$$

also follows an exponential Almon scheme.

3.3.2 Mixed-frequency VAR

Mixed-frequency VAR (MF-VAR) models extend the standard VAR framework to handle situations where high-frequency (HF) variables are observed m times within each low-frequency (LF) period. To construct a system that jointly models HF and LF variables without temporal aggregation, we stack the m HF observations within period t together with the LF variable observed at the end of the period. This imposes an intertemporal alignment constraint: the HF observations $x_{t+(j/m)}$ for $j = 0, \ldots, m-1$ all correspond to the same calendar period as the LF value y_t.

$$\begin{pmatrix} x_t \\ \vdots \\ x_{t+(m-1)/m} \\ y_{t+1} \end{pmatrix} = C_0 + \sum_{k=1}^{K_{\max}} C_k \begin{pmatrix} x_{t-k} \\ \vdots \\ x_{t-k+(m-1)/m} \\ y_{t+1-k} \end{pmatrix} + \begin{pmatrix} \varepsilon_t^1 \\ \vdots \\ \varepsilon_t^m \\ \varepsilon_t^y \end{pmatrix}. \tag{3.34}$$

where K_{max} is the maximum lag order in K_x and K_y, C_0 is constant, and C_k is $(m+1) \times (m+1)$ matrix measures within-period time series dependency.
A particular row of the MF-VAR represents HF processes predicted by past HF and LF series and vice versa. So We can connect MF-VAR with an unrestricted MIDAS by extracting the last row of it, we have

$$y_{t+1} = \beta_0 + \sum_{k=0}^{K_y} \beta_k y_{t-k} + \sum_{j=1}^{m(K_x+1)-1} \gamma_j x_{t+1-j/m} + \varepsilon_t^y \tag{3.35}$$

In the summation, m denotes the number of high-frequency observations per low-frequency period, and K_x is the maximum lag order of the high-frequency variable in the MF-VAR. The upper bound $m(K_x + 1) - 1$ therefore corresponds to the total number of high-frequency lags implied when unfolding the K_x lagged vectors of dimension m into a single regression equation.

CHAPTER 4

Data source

4.1 Sentiment

The sentiment indicators are extracted from RBA's monthly statement from Dec 2007 to Dec 2021. Since BERT can only process 512 tokens at max, and RBA's statements are well-structured, the entire statement is split into three parts: first two paragraphs (which is mostly about global economic trends and conditions on global financial markets), middle paragraphs (which is mostly about Australian economic trends), and last two paragraphs (which is mostly about cash rate decisions in nearly future and expectation of inflation). One example of RBA's statement in Dec 2020 is shown below:

First part (Global economy): *At its meeting today, the Board decided to maintain the current policy settings, including the targets of 10 basis points for the cash rate and the yield on 3-year Australian Government bonds, as well as the parameters of the Term Funding Facility and the government bond purchase program. Globally, the news has been mixed recently. On the one hand, infection rates have risen sharply in Europe and the United States and the recoveries in these economies have lost momentum. On the other hand, there has been positive news on the vaccine front, which should support the recovery of the global economy. The recovery is also dependent on ongoing support from both fiscal and monetary policy. Hours worked in most countries remain noticeably below pre-pandemic levels and inflation is low and below central bank targets.*

Last part (Forward guidance): *Given the outlook for both employment and inflation, monetary and fiscal support will be required for some time. For its part, the Board will not increase the cash rate until actual inflation is sustainably within the 2 to 3 per cent target range. For this to occur, wages growth will have to be materially higher than it is currently. This will require significant gains in employment and a return to a tight labour market. Given the outlook, the Board is not expecting to increase the cash rate for at least 3 years. The Board will keep the size of the bond purchase program under review, particularly in light of the evolving outlook for jobs and inflation. The Board is prepared to do more if necessary.*

The topic of these three parts can be summarized as global economy, domestic economy, and forward guidance. So three dimensions of monetary policy represented by the sentiment score of three parts are used as indicators here.

Evidence from my preliminary analysis Chen (2022), as well as insights from the literature on central bank communication, suggests that forward guidance sentiment tends to be the most important for predicting macroeconomic variables. This is intuitive: forward guidance directly conveys the RBA's expectations about future inflation, interest rates, and economic conditions, and therefore embeds information about the path of monetary policy that markets and households react to most strongly. Sentiment in this section is typically forward-looking and therefore aligns closely with the forecasting objective of this thesis.

Nevertheless, it remains meaningful to include all three sentiment indicators. While forward guidance captures the most explicit policy signal, both global and domestic economic sentiment provide important contextual information about the RBA's assessment of economic conditions and risks. These two dimensions shape the policy narrative by offering explanations for the RBA's decisions and by signalling how external or domestic shocks are evolving over time. Including all three indicators allows the forecasting model to capture a richer structure of monetary policy communication, one that reflects both the rationale behind policy actions (global and domestic sentiment) and the expected trajectory of policy (forward guidance sentiment). Empirically, this comprehensive approach ensures that the model leverages the full informational content embedded in the RBA's statements, rather than relying solely on a single dimension of communication.

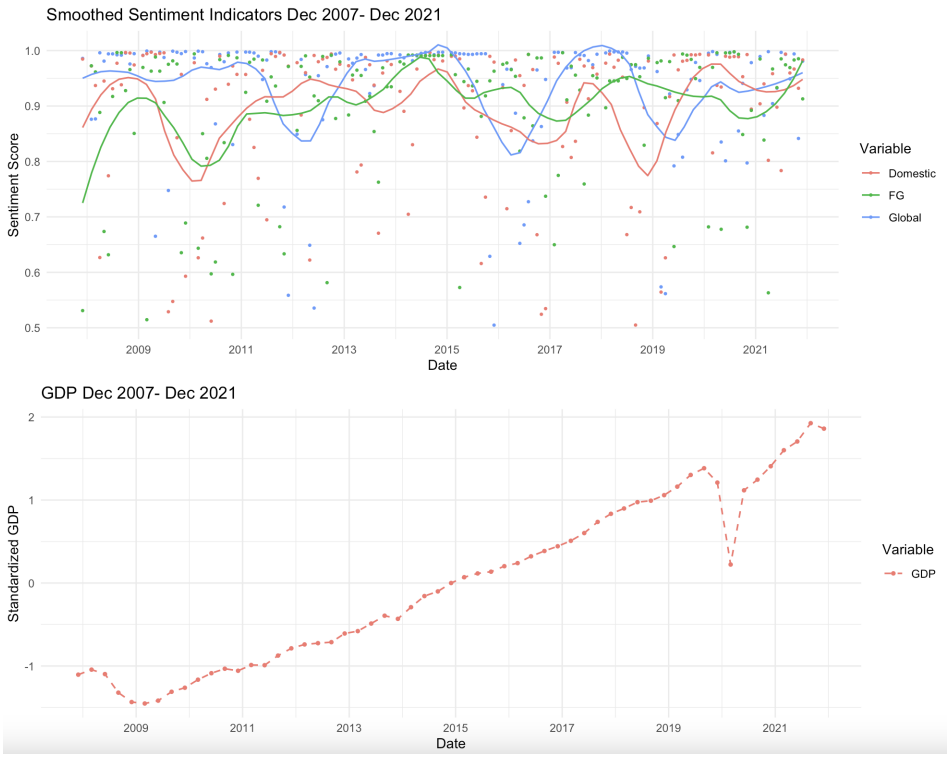

Figure 4.1: Time Series Plot of Sentiment and GDP (2017-2021)

4.2 Macroeconomic data

Australia GDP as the main macroeconomic indicator is collected from ABS. Figure 4.1 shows the time plot of three sentiment and standardized GDP. A key observation is that GDP exhibits substantial fluctuations, particularly during the COVID-19 period (early 2020), where it shows a sharp decline followed by a strong recovery. However, the sentiment indicators stay relatively stable, with some variations but no clear trend that strongly mirrors GDP's movements. Notably, while a decreasing in sentiment indicators is visible around early 2020, it does not exhibit the same magnitude of volatility as GDP. This suggests that sentiment indicators, while possibly reflecting some economic disruptions, may not have fully captured the extreme macroeconomic shocks during the pandemic period. The weak association between GDP and sentiment in this plot raises the question of whether additional sentiment measures, or interaction effects with other macroeconomic variables could enhance the forecasting relationship. Therefore, this study will also compare whether sentiment adds forecasting power above including high-frequency financial variables like asset prices.

Panagiotelis et al. (2019) evaluated the empirical performance of the competing forecasting methods with nested subsets of predictors, the full data set contains 151 variables. They found substantial gains in the accuracy of forecasts of the key macroeconomic variables of interest after increasing the information set from 3 to 13 predictors. Since GDP are used as the main indicator and forecasting aim in this paper, and cash rate are fully reflected in RBA's statements so can be captured by sentiments. This paper uses remaining 11 most important variables in macroeconomic forecasting (see Table 2). To avoid high dimensions in the dynamic factor model, these 11 variables are summarized into one common factor and used as an external indicator.

Table 4.1: Principal Component Analysis of key macroeconomic variables

	Standard Deviation	**Proportion of Variance**	**Cumulative Proportion**
1	2.6237	0.8605	0.8605
2	0.70947	0.06292	0.92340
3	0.58538	0.04283	0.96623
4	0.4656	0.0271	0.9933
5	0.18187	0.00413	0.99747
6	0.11207	0.00157	0.99904
7	0.07601	0.00072	0.99976
8	0.04379	0.00024	1.00000

The Principal Component Analysis (PCA) results are shown in Table 4.1, it indicate that PC1 alone explains approximately 86.05% of the total variance, making it the most dominant component in capturing the underlying structure of the dataset. The cumulative variance explained by PC1 and PC2 together reaches 92.34%, suggesting that nearly all the essential information in the original dataset can be represented by just these two components. The remaining components contribute minimally, with

PC3 adding only 4.28% and PC4 just 2.71%, reinforcing the idea that higher-order components capture only marginal variations.

In this research, PC1 is extracted as an external indicator in forecasting models. Moreover, PCA is implemented using an expanding window approach, which ensures that the principal components are continuously updated as new data becomes available. This method offers several advantages: it allows the model to adapt to structural changes in the data, captures evolving relationships between variables, and mitigates potential issues arising from parameter instability over time. By dynamically updating PC1, this approach enhances the robustness of forecasting by ensuring that the extracted common factor remains representative of the most recent economic conditions, making it a more effective predictor in a dynamic macroeconomic environment.

CHAPTER 5

Results analysis

This section presents a detailed analysis of the forecasting performance of the six models used in this study under expanding window: ARIMA(1,1,0) as the naive benchmark, Auto ARIMA, Auto ARIMA with the common factor (PC1) of 11 most important variables as the external indicator, MIDAS, Dynamic Factor model (DFM) estimated by Kalman Filter, and MF-VAR. Table 5.1 presents the forecast accuracy across forecasting horizons from 1 quarter to 4 quarters, incorporating results from the Model Confidence Set (MCS) test. The MCS test identifies models that significantly outperform others at each horizon, indicated by boxed values in the table, while the best-performing models are highlighted in bold. The training/test split is 80/20.

Auto ARIMA significantly improves upon ARIMA(1,1,0), particularly at the short forecasting horizon. At horizon 1, its RMSE (0.6312) is approximately 48.35% of the benchmark model's RMSE, indicating strong short-term predictive power. However, its relative RMSE declines slightly as the horizon increases, suggesting diminishing predictive power relative to benchmark in longer-term. Including the common factor as an external regressor in the Auto ARIMA model further improves forecasting accuracy. This model maintains a lower RMSE across all horizons than the benchmark and plain Auto ARIMA. As the horizon increases, Auto ARIMA with PC1 consistently outperforms Auto ARIMA, indicating that incorporating common factors from macroeconomic variables helps sustain predictive accuracy over longer periods.

For mixed-frequency models, MIDAS demonstrates a notable ability to capture high-frequency information and improve forecasting accuracy. At horizon 1, its relative RMSE (0.4452) is lower than all traditional time-series models and is highlighted as a statistically superior model under the MCS test. As the forecast horizon extends, MIDAS maintains stable performance, consistently outperforming benchmark models. DFM emerges as the best-performing model across all forecasting horizons, as expected. At horizon 1, its relative RMSE (0.3544) is the lowest among all models, representing only 35.44% of the benchmark RMSE. It sustains superior accuracy across all horizons, with its relative RMSE remaining lower than all competing models, particularly at horizon 4 (0.3391). The MCS test confirms DFM's dominance at all horizons, reinforcing its ability to dynamically incorporate high-frequency predictors through state-space modeling. This result indicates the effectiveness of state-space modeling and the ability of the Dynamic factor model to dynamically incorporate higher frequency predictors. The MF-VAR model performs well but does not significantly outperform DFM or even Auto ARIMA with PC1 especially in longer horizons. While its RMSE at horizon 1

(0.4436) is competitive, it does not achieve statistical superiority under the MCS test. Additionally, it experiences greater deterioration at longer horizons. At horizon 4, its relative RMSE (0.5985) is significantly higher than that of MIDAS and DFM, suggesting that it may struggle with capturing persistent predictive patterns over extended periods.

The findings reveal that models leveraging high-frequency information, such as MIDAS and DFM, tend to outperform traditional time-series models, particularly for short-term forecasts. DFM emerges as the most robust model across all horizons, maintaining the lowest RMSE. MIDAS also proves to be a competitive alternative model, particularly for mid-range forecasting. In contrast, Auto ARIMA can provide a relative stronger short-term performance than benchmark but lacks consistency in longer-term horizons. Also, including external indicators such as common factor from macroeconomic variables can enhance forecasting performance, as demonstrated by Auto ARIMA with PC1, but still dominated by mixed-frequency models.

Overall, DFM demonstrates the strongest predictive performance, making it the most suitable choice for macroeconomic forecasting in this study. MIDAS also provides a competitive alternative, particularly for long-term forecasting. The results show the importance of integrating higher-frequency sentiment predictors into forecasting models to improve predictive accuracy, aligning with the broader objective of enhancing macroeconomic forecasting with dynamic factor models in a state-space framework.

Table 5.1: Forecasting accuracy for forecasting horizon from 1 to 4 quarter.

Horizon	**h=1**	**h=2**	**h=3**	**h=4**
ARIMA(1,1,0)	1.3055	1.0488	1.4521	1.8480
Auto ARIMA	0.5036	0.9524	0.8463	0.7262
Auto ARIMA with PC1	0.4835	0.9095	0.7974	0.6732
MIDAS	0.4452	0.6005	0.4735	0.3891
DFM	**0.3544**	**0.4955**	**0.3884**	**0.3391**
MF-VAR	0.4436	0.8969	0.7391	0.5985

Each entry in the first row shows the RMSE of ARIMA(1,1,0) as the benchmark model. The other lines shows the relative RMSE to the benchmark. Bold entries indicate the lowest error measure achieved by the competing models in each horizon. Entries in rectangle represent the model(s) significantly dominate others under MCS test for each horizon.

As MIDAS and MF-VAR were selected as 'winner' above, I next added high-frequency financial variables into these two models in Table 5.2, to evaluate whether sentiment indicators improve forecasting performance beyond the inclusion of high-frequency financial variables, specifically the S&P/ASX 200 Adjusted Close index. Across all forecasting horizons, mixed-frequency models significantly outperform the Auto ARIMA model, indicates that both kind of high-frequency information can provide forecasting power above lower-frequency factors. Among mixed-frequency models, MIDAS-S&ASX

Table 5.2: Forecast accuracy compared with ASX index

Horizon	**h=1**	**h=2**	**h=3**	**h=4**
Auto ARIMA with PC1	0.4835	0.9095	0.7974	0.6732
MIDAS-S	0.4452	0.6005	0.4735	0.3891
MIDAS-ASX	0.4231	0.5593	0.4204	0.3399
MIDAS-S&ASX	0.4233	0.5601	0.4255	**0.3306**
DFM-S	0.3544	0.4955	**0.3884**	0.3391
DFM-ASX	0.3518	0.5058	0.4119	0.3533
DFM-S&ASX	**0.3048**	**0.4815**	0.3969	0.3418

Each entry shows the relative RMSE to the benchmark ARIMA(1,1,0). Bold entries indicate the lowest error measure achieved by the competing models in each horizon. MIDAS-S includes sentiment, MIDAS-ASX includes ASX index only, MIDAS-S&ASX includes both sentiment and ASX index. Entries in rectangle represent the model(s) significantly dominate others under MCS test for each horizon.

generally outperforms MIDAS-S and often performs closely to MIDAS-ASX, demonstrating that combining sentiment indicators with financial market data can improve predictive accuracy. For example, at horizon 1, the relative RMSE for MIDAS-S is 0.4452, while MIDAS-S&ASX achieves 0.4233, supporting the idea that integrating both sentiment and financial variables yields superior forecasts. This trend continues for longer horizons, with MIDAS-S&ASX maintaining lower relative RMSE values than MIDAS with sentiment alone. Notably, models that include only sentiment (MIDAS-S) tend to perform worse than those incorporating financial variables. For instance, at horizon 4, the RMSE for MIDAS-S is 0.3891, whereas MIDAS-ASX achieves 0.3399, highlighting that financial variables alone provide stronger predictive power than sentiment alone. However, MIDAS-S&ASX improves further to 0.3306, suggesting that while sentiment alone may not be the most effective predictor, it does provide incremental value when combined with financial market data.

Additionally, comparing among three Dynamic factor models, the results give a similar structure with MIDAS models. DFM-ASX (0.3518 at horizon 1) performs better than DFM-S in short horizon but not longer ones, reinforcing that financial data can enhance forecasts in short-term. DFM-S&ASX (0.3048 at horizon 1, 0.4815 at horizon 2) further improves forecast accuracy, and dominates all other models in short horizons. These results align with the time plot analysis in Figure 4.1, suggesting that in short horizons, sentiment indicators alone may not fully capture economic fluctuations, particularly during extreme events such as COVID-19. The weak association between sentiment and GDP in the time plot is reflected in its relatively weaker forecasting performance alone. However, when sentiment is combined with high-frequency financial variables, it provides additional predictive power, possibly by capturing non-financial economic signals that asset prices alone might not reflect, highlighting the complementary nature of sentiment indicators and financial market data. The results emphasize the importance of incorporating both sentiment and financial market information to enhance macroeconomic forecasting accuracy, supporting a multi-source data integration approach for

improved predictive performance.

A useful benchmark for assessing the effectiveness of the proposed methodology is the Aruoba-Diebold-Scotti (ADS) Business Conditions Index, which is commonly used in the U.S. for tracking real-time economic conditions Aruoba et al. (2009). While the ADS Index relies on Dynamic factor model applied to macroeconomic indicators such as employment, industrial production, and income, my approach integrates high-frequency sentiment indicators and asset prices into the forecasting model. This comparison raises key conceptual differences. Firstly, ADS Index captures broad macroeconomic conditions, but its frequency is constrained by the availability of underlying data. In addition, sentiment and asset price-based forecasting reacts more quickly to economic changes, as financial markets and news sentiment adjust in real-time. Moreover, ADS relies on structural macroeconomic relationships, while my approach captures market expectations and behavioral responses, which may be leading indicators of economic changes.

Table 5.3: Forecast accuracy pre- and post-COVID (relative RMSE to ARIMA(1,1,0))

Panel A: Pre-COVID (up to 2020Q1)

Model	**h=1**	**h=2**	**h=3**	**h=4**
Auto ARIMA with PC1	0.4481	0.9310	0.8495	0.7896
MIDAS-S	0.4175	0.5845	0.4473	0.4044
MIDAS-ASX	0.4260	0.4803	0.3564	0.3275
MIDAS-S&ASX	0.4231	0.5349	**0.3098**	**0.2461**
DFM-S	0.3439	0.4904	0.3241	0.2960
DFM-ASX	0.3397	0.5169	0.3963	0.3145
DFM-S&ASX	**0.2875**	**0.4566**	0.3782	0.3274

Panel B: Post-COVID (after 2020Q1)

Model	**h=1**	**h=2**	**h=3**	**h=4**
Auto ARIMA with PC1	0.5017	0.9526	0.8439	0.7254
MIDAS-S	0.3733	0.5411	0.4679	0.3584
MIDAS-ASX	0.4714	0.4460	0.4259	0.3901
MIDAS-S&ASX	0.4445	0.3909	**0.3336**	0.3626
DFM-S	0.3789	0.4074	0.3664	0.3158
DFM-ASX	0.3846	0.3971	0.3729	0.3471
DFM-S&ASX	**0.3310**	**0.3833**	0.3441	**0.3110**

Notes: Each entry reports relative RMSE compared with benchmark ARIMA(1,1,0). Bold entries indicate lowest RMSE for each horizon. Shaded cells indicate models included in the 10% Model Confidence Set (MCS). Panel A uses data up to 2020Q1; Panel B uses the post-COVID sample.

The results in Table 5.3 reveal a clear distinction between pre- and post-COVID forecast performance across all models. In the pre-COVID period, models that incorporate both sentiment and financial information consistently outperform their competitors. In particular, the DFM-S&ASX model delivers the lowest relative RMSE across short forecast horizons, indicating that the combination of structured textual sentiment and

high-frequency financial indicators provides valuable leading information about macroeconomic conditions in a stable environment. MIDAS-based specifications also perform well, especially at longer horizons, suggesting that the MIDAS framework effectively exploits high-frequency predictors when the underlying economic relationships are stable and well behaved. Overall, the pre-COVID results highlight the importance of combining mixed-frequency predictors and more flexible modelling structures when forecasting macroeconomic variables such as GDP.

After the onset of COVID-19, forecast performance deteriorates for most models, reflecting the extreme volatility and structural breaks introduced during the pandemic period. Nevertheless, several models continue to exhibit relatively strong performance. For instance, DFM-S&ASX remains one of the best-performing specifications at horizons h = 1, 2, and 4, while MIDAS-S&ASX provides the best performance at h = 3. These results suggest that incorporating the ASX index as an immediately responsive financial market indicator can help models adjust more rapidly to sudden economic disruptions. Sentiment-only specifications generally underperform combinations that include ASX, indicating that financial markets contain unique information during crisis periods that is not fully captured in policy communication. Overall, the post-COVID results reinforce the value of integrating real-time financial indicators alongside sentiment measures, while also illustrating the challenges faced by traditional time-series models in periods of heightened uncertainty and structural change.

Overall, both approaches have complementary strengths. The ADS Index is well-suited for tracking business cycles, while high-frequency sentiment and asset prices may provide superior short-term forecasting capabilities, particularly during periods of rapid economic change. Since no direct Australian equivalent of the ADS Index exists, our research presents an alternative approach that integrates real-time financial and sentiment-driven data into GDP forecasting. Future work could explore combining ADS-style macroeconomic indicators with sentiment and financial data to further improve macroeconomic forecasting accuracy.

CHAPTER 6

Conclusion

This study investigates whether incorporating sentiment extracted from central bank communication as a high-frequency indicator in mixed-frequency models can enhance macroeconomic forecasting performance. Additionally, it evaluates the comparative effectiveness of different forecasting models, including MIDAS, Kalman Filter-based Dynamic Factor Models (DFM), and mixed-frequency vector autoregression (MF-VAR). The results provide strong evidence that high-frequency sentiment indicators contribute valuable information to GDP forecasting, particularly when combined with financial market variables.

The findings indicate that MIDAS and DFM models incorporating sentiment (MIDAS-S, DFM-S) generally improve forecast accuracy over traditional models. However, sentiment alone does not consistently outperform financial variables such as the ASX index especially in short horizons. In short term, some information from RBA communication has already reflected in financial variables. Therefore, the best-performing models combine both sentiment and asset price information, as evidenced by the superior accuracy of MIDAS-S&ASX and DFM-S&ASX, at both short and long horizons. The MCS test further confirms the statistical significance of these models, reinforcing that combining multiple high-frequency sources enhances predictive power. Compared to past research, which has relied on traditional macroeconomic indicators and low-frequency models, this study highlights the benefits of including high-frequency textual data with financial data to improve macroeconomic forecasting.

Despite its contributions, this study has several limitations that open the door to further research. First, the empirical analysis does not incorporate real-time or "nowcasting" macroeconomic indicators, such as weekly labour market data, electricity consumption, mobility indices, or payment-system data. These increasingly common high-frequency indicators may capture business-cycle dynamics not fully reflected in sentiment or financial markets, and integrating them into a mixed-frequency framework could substantially improve short-horizon forecasts.

Second, the study focuses exclusively on sentiment extracted from central bank communication. While RBA statements are an important and policy-relevant source of information, additional textual sources such as economic news articles could provide a more comprehensive and diversified measure of real-time sentiment. Central bank communication is measured monthly and tends to be highly curated; in contrast, other media and market commentary respond instantaneously to economic developments and may reveal risk perceptions or uncertainty not present in official statements. Third, the fore-

casting framework employs a linear dynamic factor model with Gaussian assumptions. Modern NLP techniques and mixed-frequency machine-learning models (e.g., nonlinear DFMs, or random forests for mixed frequencies) could uncover nonlinear relationships between sentiment, financial conditions, and macroeconomic outcomes. Exploring whether machine-learning alternatives can outperform the linear DFM would be a natural extension. Fourth, this thesis restricts attention to a univariate forecasting target

(GDP). However, the joint modelling of multiple macroeconomic variables—including inflation, consumption, investment, and labour market indicators—could allow for a richer assessment of the channels through which sentiment and financial markets influence the economy. Extending the framework to a multivariate setting may also improve forecast accuracy via cross-series information sharing. Finally, this study abstracts from

real-time data revisions. RBA statements are published in real time, while macroeconomic variables undergo potentially large revisions. A potential avenue for future work would incorporate real-time data vintages and consider how sentiment interacts with revised vs. unrevised macro datasets, especially given evidence that central bank communication often responds to information not yet reflected in official statistics.

Overall, this research makes several key contributions to the existing literature on macroeconomic forecasting:

1. **Innovative use of high-frequency sentiment indicators with financial indicators**: As mentioned above, unlike past studies that primarily use structured macroeconomic indicators, this study demonstrates that using sentiment extracted from central bank communication and high-frequency financial variable together can enhance forecasting accuracy when integrated into mixed-frequency models.

2. **Comparison of alternative mixed-frequency models**: By evaluating MIDAS, DFM, and MF-VAR models, this study firstly provides a comprehensive assessment of their relative forecasting performance, extending some prior papers that often focus on a single model class. In addition, some other previous related papers compared different models, but most of them just used macroeconomic variables. My study makes the comparison when textual data is including as well, which can give a more convincing comparison result.

3. **Empirical validation using Australian data**: Most existing research on high-frequency forecasting focuses on U.S. and European economies. This study contributes to the Australian macroeconomic forecasting literature, offering new insights into how sentiment and financial variables interact in a mixed-frequency framework in Australia economy.

Overall, this study provides strong evidence that incorporating sentiment indicators from central bank communication into mixed-frequency models improves GDP forecasting accuracy, particularly when combined with financial market data. These findings shows the value of high-frequency textual analysis in macroeconomic forecasting and suggest potential directions for future research, including the integration of real-time macroeconomic data and a broader range of sentiment sources to enhance forecasting precision.

CHAPTER 7

Appendix

Table 1: Transformation code

Transformation (T)	Expression
1	w_t
2	$w_t - w_{t-1}$
3	$(w_t - w_{t-1}) - (w_{t-1} - w_{t-2})$
4	$\ln(w_t)$
5	$\ln(w_t/w_{t-1})$
6	$\ln(w_t/w_{t-1}) - \ln(w_{t-1}/w_{t-2})$

w_t denotes an observed variable in levels. Transformation (T) denotes the transformation implemented to achieve stationarity: 1 = no transformation; 2 = first difference; 3 = second difference; 4 = log; 5 = first difference of logged variables; and 6 = second difference of logged variables.

Table 2: Macroeconomic variables used in PC1

Name	T	Description
CPI-ALL	6	Index Numbers; All groups CPI
IP_TotalInd	5	Total industrial industries; Index
Emp_TotalPer	5	Employed – total; Persons
Hstarts_PDA	5	Private dwelling approvals
COMM	6	Index of commodity prices; All items; AUD, Index, 2013/14=100
Exports	5	Exports of goods and services
Imports	5	Imports of goods and services
M1	6	M1
Credit_Total	6	Credit; Total
SP ASX AllOrds	5	S&P ASX AllOrds adjusted closing prices

Bibliography

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

Aruoba, S.B., Diebold, F.X., Scotti, C., 2009. Real-Time Measurement of Business Conditions. Technical Report 09-19. Federal Reserve Bank of Philadelphia. URL: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/ads.

Assimakis, N., Adam, M., 2009. Steady state kalman filter for periodic models: A new approach. International Journal of Contemporary Mathematical Sciences 4, 201–218.

Bai, J., Ghysels, E., Wright, J.H., 2013. State space models and MIDAS regressions. Econometric Reviews 32, 779–813.

Barbaglia, L., Consoli, S., Manzan, S., 2023. Forecasting with economic news. Journal of Business & Economic Statistics 41, 708–719.

Barbaglia, L., Consoli, S., Manzan, S., 2024. Forecasting GDP in Europe with textual data. Journal of Applied Econometrics 39, 338–355.

Benchimol, J., Kazinnik, S., Saadon, Y., 2025. Federal Reserve Communication and the COVID-19 Pandemic. The Manchester School .

Bernanke, B.S., Boivin, J., Eliasz, P., 2005. Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. The Quarterly journal of economics 120, 387–422.

Bholat, D., Broughton, N., Parker, A., Ter Meer, J., Walczak, E., 2018. Enhancing central bank communications with behavioural insights .

Blinder, A.S., 2009. Talking about monetary policy: the virtues (and vice?) of central bank communication .

Carriero, A., Galvao, A.B., Kapetanios, G., 2019. A comprehensive evaluation of macroeconomic forecasting methods. International Journal of Forecasting 35, 1226–1239.

Chen, M., 2022. A computational linguistics approach to study the macroeconomic effects of central bank communication: The case of the rba. Master thesis University of Sydney.

Ellingsen, J., Larsen, V.H., Thorsrud, L.A., 2022. News media versus FRED-MD for macroeconomic forecasting. Journal of Applied Econometrics 37, 63–81.

Fildes, R., Stekler, H., 2002. The state of macroeconomic forecasting. Journal of macroeconomics 24, 435–468.

Foroni, C., Marcellino, M., Schumacher, C., 2015. Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. Journal of the Royal Statistical Society Series A: Statistics in Society 178, 57–82.

Ghysels, E., Santa-Clara, P., Valkanov, R., 2004. The MIDAS touch: Mixed data sampling regression models .

Haldane, A., McMahon, M., 2018. Central bank communications and the general public, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203. pp. 578–583.

Hansen, S., McMahon, M., 2016. Shocking language: Understanding the macroeconomic effects of central bank communication. Journal of International Economics 99, S114–S133.

He, C., 2021. Monetary policy, equity markets and the information effect. Reserve Bank of Australia Sydney.

Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., Kapadia, S., 2022. Making text count: economic forecasting using newspaper text. Journal of Applied Econometrics 37, 896–919.

Kuzin, V., Marcellino, M., Schumacher, C., 2011. MIDAS vs. mixed-frequency VAR: Nowcasting GDP in the euro area. International Journal of Forecasting 27, 529–542.

Lowe, P., 2020. Responding to the Economic and Financial Impact of COVID-19. Speech at the Reserve Bank of Australia, Sydney 19.

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

Mariano, R.S., Murasawa, Y., 2010. A coincident index, common factors, and monthly real GDP. Oxford Bulletin of Economics and Statistics 72, 27–46.

Okuneva, M., Hauber, P., Carstensen, K., Bär, J., 2024. Nowcasting German GDP with Text Data. Technical Report. CESifo Working Paper.

Panagiotelis, A., Athanasopoulos, G., Hyndman, R.J., Jiang, B., Vahid, F., 2019. Macroeconomic forecasting for Australia using a large number of predictors. International Journal of Forecasting 35, 616–633.

Schorfheide, F., Song, D., 2015. Real-time forecasting with a mixed-frequency VAR. Journal of Business & Economic Statistics 33, 366–380.

Shapiro, A.H., Sudhof, M., Wilson, D.J., 2022. Measuring news sentiment. Journal of econometrics 228, 221–243.

Stock, J.H., Watson, M.W., 2002. Macroeconomic forecasting using diffusion indexes. Journal of Business & Economic Statistics 20, 147–162.

Stock, J.H., Watson, M.W., 2011. Dynamic factor models .

Ter Ellen, S., Larsen, V.H., Thorsrud, L.A., 2022. Narrative monetary policy surprises and the media. Journal of Money, Credit and Banking 54, 1525–1549.