Abstract

Macroeconomic forecasts enable the policy-makers to foresee the future economic trends and take prompt measures to ensure longer economic growth and quicker economic recovery. Accurate and timely macroeconomic forecasts may also help the enterprises to make better long-term business strategies. Social media has a swift and sensitive response to economic dynamics through online news reports, interviews and individual comments. Economic Indices extracted from social media data are more immediate and comprehensive, but lack of stability and credibility in empirical studies. This research proposed a mix frequency modelling approach to incorporate only the recent high frequency part of social media data in traditional econometrics based macroeconomic forecasting with support of a multisource based macroeconomic forecast system. A mixed data sampling (MIDAS) model is constructed and an empirical evaluation is presented to show how to incorporate Google search queries into Chinese CPI forecasting. The empirical results indicate a satisfactory improvement in forecasting performance. The multisource modelling and forecasting framework offers a practical and implementable solution for involving social media data sources into macroeconomic forecasting systems. This research contributes to future development of decision support systems for governmental policy-making and enterprises’ operational decisions in the Big Data era.

Keywords: Social media data, Macroeconomic forecast, Mixed data sampling, Consumer Price Index.
1 INTRODUCTION

Macroeconomic forecast has been of great significance for the macroeconomic policy making since the past decades. Timely and accurate forecasts of the important macroeconomic indicators can aid the government to control the gross domestic production, inflation levels, unemployment rates and international payments balance, all of which are the concerns of the macroeconomic policy. Thus, the task of macroeconomic forecast draws attention of governments, academics and enterprises. For the government officials, the timely and accurately forecasting can help to support the related macroeconomic policymaking. For the enterprises, especially securities companies, they are particularly concerned with the future macroeconomic trends which can instruct their investment and development strategies. For academic researchers, forecasting the main macroeconomic indicators to help government and enterprises to make decisions has the challenge of modelling the complex economic system, managing the uncertainty and handling large volumes of data from multiple sources.

The social media data from online news reports, interviews and individual comments is a more immediate and comprehensive alternative source for macroeconomic forecasting. It is created and publicly available in real time and contains information about almost every perspectives of life and business. For example, many well-received news websites, such as ‘Sina Finance’, ‘He Xun’, ’Yahoo Finance’, 'The Wall Street Journal', update news every day. The news contains abundant information that reflects the judgement or the expectation the public hold. There are some social media platforms, such as forums, micro-blogs, can also provide rich social media data reflecting the macroeconomic trends. In addition, the public tend to show their concerns on some specific issues via the Internet. The search queries based on related keywords reflect the expectation on the macroeconomic issues (Askitas and Zimmermann 2009; Guzman 2011; McLaren and Shanbhogue 2011). However, since social media data is often ill-structured and text-based, it will be a great challenge in incorporating it into the macroeconomic forecast. Existing researches have tried to employ machine learning method to extract meaningful signals to construct the econometric models (Choi and Varian 2009; Messinger et al. 2009; Choi and Varian 2012). The econometric modelling approaches, such as Autoregressive Integrated Moving Average (ARIMA), Autoregressive Distributed Lag (ADL), Error Correct Model (ECM), Dynamic Factor models (Stock and Watson 2005; Forni et al. 2005) etc. remain the most used ones in empirical forecasts for the practice. But the traditional econometric models always lacks of considering the high frequency part of the data, which is just the specical characteristics of the soical media data. Under the circumstance, a mixed data sampling (MIDAS) approach is introduced to solve the problem. The MIDAS approach can combine the high frequency social media data and the traditional statistical data, and forecast the future values of the macroeconomic indicators.

The research is motivated by two research gaps. Firstly, the recent macroeconomic forecast seldom employs the timely updated social media data, which is usually at higher frequency compared to the traditional economic data. This research provides a practical solution for government and enterprise on how to apply the social media data into the macroeconomic forecast. Secondly, this paper introduces a MIDAS approach, which incorporates the high frequency part of social media data to forecast macroeconomic indicators. It is quite different from most existing papers in social media data aggregation methods, which convert the social media data into the relative low frequency time series. The MIDAS method makes use of as much information contained by the high frequency part of the social media data as possible. Moreover, the empirical results prove that the MIDAS approach performs better in forecasting the China’s inflation level.

This paper attempts to incorporate the social media data into the macroeconomic forecast with econometric modelling techniques to improve the forecast precision rate. A mixed data sampling (MIDAS) approach is introduced to solve the data modelling of different frequencies. The importance of each day to the whole month is various, so the average, the maximum or the minimum could not represent the whole forecast information. Besides, a systematic macroeconomic forecast framework, which integrates multisource data, including low-frequency statistical data and high-frequency social
Social media data, utilizes multiple econometric models, is proposed. According to the proposed framework, an empirical study on CPI forecast utilizing the social media data from the Google search engine, with the help of mixed frequency modelling is presented. The forecast results indicate that model incorporating the social media data improves the forecasting accuracy compared with other models without it.

The rest of the paper is organized as follows: Section 2 reviews the relevant literature. Section 3 illustrates the mixed frequency models. Section 4 describes the design of multisource macroeconomic forecast system. Section 5 provides the empirical study on CPI forecast. Finally, Section 6 draws conclusions and gives directions for the future work.

2 RESEARCH BACKGROUND

2.1 Social Media Data Retrieval and Analysis

Social media data can be crawled from web portals, forums, social network websites, weblogs, and micro-blogs and so on (Wang et al. 2007; Parameswaran et al. 2007). The widespread use of the Internet by both business and consumers has led to the creation of a potentially useful social media data. It is easily available and updated timely at high frequency (Asur and Huberman 2010). In this paper, the social media data is classified into three types according to different sources. The first type is the related news reports crawled from the authoritative news websites. The second type is comments and reviews extracted from some platforms, including forums, micro-blogs and weblogs and so on. And the last type is the search queries generated from main search engines.

Social media data retrieval and analysis methods vary with the different data types. For the first type, the traditional information retrieval and text mining analysis can be used to analyse it, since the data is mainly in the forms of texts. For example, Probability models had been used for information retrieval for decades (Maron and Kuhns 1960). And the first application was the “standard probability model” (Robertson and Sparck 1976). Based on the previous work, hidden Markov models were developed for retrieval. The general idea of these methods is to rank the documents according to the probability of a document’s relevance to other sets of documents given a user-generated query. Vector Space Model (VSM) is also a standard tool for data retrieval. TF-IDF is used to compute the term frequency and inverse document frequency that reflects the importance of different words in the English language (Frakes 1992). Latent semantic indexing (LSI) and principal component analysis (PCA) are two dimensional reduction methods used in very large data sets (Berry and Castellanos 2004). After the data retrieval, data analysis needs to be utilized to obtain the time series data. For example, regression analysis, factor analysis, ID3 and C4.5 (Bai et al. 2007; Jain et al. 2012; Carpineto and Romano 2012). For the second type of social media data, sentiment analysis can be used to analyze it. Sentiment analysis is concerned with analyzing direction-based text, determines whether a text is objective or subjective (Dang et al. 2010; Chen et al. 2012). It can be considered as a common two-class problem that involves classifying sentiments as positive or negative. A text’s sentiment is determined by the sentiments of a group of words or phrases appearing in the text. And the machine learning approach and semantic orientation approach are the common approaches to deal with the sentiment analysis of the social media data.

For the last type of social media data, some search engines provide the search query volume for download based on the keywords defined by the users. Many researchers have used the service provided by Google to extract the social media data (Ettredge et al. 2005; Goel et al. 2010; Choi and Varian 2012). The search terms defined by the users could represent their concerns on the macroeconomic trends, which are recorded by the search engines. The goal of social media data analysis is to obtain the time series data used for correlation analysis and forecasting. After the data retrieval and analysis, the high frequency weekly or daily time series is acquired.
2.2 Applications in Forecasting

As social media data is timely and high frequency compared to statistical data, some researchers have attempted to add different kinds of social media data to help their analysis and forecast in the economic fields. The following parts will illustrate the applications of social media data in forecasting from three perspectives. Firstly, important news websites provide social media data useful for macroeconomic forecast (Li and Wu 2010). For example, ‘Sina Finance’ offers the reader economic news each day. The news related to specific economic indicator, like CPI, can be crawled regularly from the websites according to related keywords. Then, data analysis is applied to extract the time series correlated with the economic indicator. Therefore, a new index extracted from the news can be constructed. Secondly, social media data generated from some social media platforms should be incorporated into the macroeconomic forecast as well. The new indicator extracted from the news websites, reviews, blogs, and forums and so on is constructed to forecast the inflation in China (Qu and Shang 2012). They defined a set of keywords based on the rise and fall of prices relevant to inflation, and then crawled a number of the search queries from many platforms. Their work demonstrated that the social media data from the general websites can help improve the traditional forecast. Thirdly, the search engines queries, like Google, Baidu, are used for extracting the social media data. The search query volumes from Google Trends have been applied to forecast U.S. unemployment rate, and the forecast precision rate is higher than other models without the social media series (Ettredge et al. 2005). More aspects of the social media data include early detection of influenza epidemics (Ginsberg et al. 2008), private consumption forecasting (Kholodilin et al. 2010; Vosen and Schmidt 2012), automobile sales forecasting (Choi and Varian 2012), and business cycle indication (Chen 2010) and so on (Swallow and Labbe 2010; Humphrey 2010). The strong correlation between the search query volumes based on the specific economic indicator and the macroeconomic trends were demonstrated in above literatures as well.

The above existing researches will be evaluated from data acquisition, pre-processing, modelling and forecasting evaluation. Firstly, the search queries are obtained according to related keywords. For example, the defined keywords were based on the research topics, traffic conditions, living conditions and the property crimes (Zhu et al. 2012). Secondly, most of the pre-processing methods are converting the weekly search queries into the monthly series by means of averaging. Thirdly, the simple models are constructed, like ARIMA and ADL models. Lastly, forecast validation methods are used to assess the validity of the search queries and the forecasting performance compared with benchmark models, and different criteria are selected to compare the forecast precision. The existing literatures applying the social media data into the macroeconomic forecast seldom consider the high frequency of the social media data, which is the speciality of the social media data compared with the statistical macroeconomic data.

2.3 Decision Support Systems based on Social Media Data

In order to make macroeconomic forecast generate efficient decision making, it is necessary to develop a decision support system based on multi source data. Actually, there are already decision support systems supporting to make decisions in China. For example, Center for Forecasting Science, China Academy of Science (CEFS), has developed the decision support systems to support PBC and National Development and Reform Committee (NDRC) with their policy making. The macroeconomic systems are based on related economic theories and utilize the statistical data from SSB or PBC. Econometric models are incorporated into the decision support systems to support the important macroeconomic indicators forecasting. It has been verified that the decision support systems are extremely useful for the decision making in China’s macroeconomic forecasting practices (Al-Othman 2006; Zhang 2009; Azadeh et al. 2012).

Social media data technologies have been used to facilitate development and management of decision making support system (DMSS) (O’Leary 2011). Dai et al. (2011) constructed a model for detecting
competitive intelligence from social media. They introduced SoMEST (Social Media Event Sentiment Timeline), a novel CI analysis framework for social media and the architecture of a natural language processing tool. The system was used to analyze the events and opinions to support decision-making. Abdul-Mageed et al. (2012) presented SAMAR, a system for subjectively sentiment analysis for Arabic social media genres.

The DSS embedded social media data is mainly on the reviews, products, business and other micro related aspects. Few systems support the macroeconomic decision systems incorporating the social media data. As is well known, macroeconomic forecast is a complex giant system with its own speciality (Chen et al. 2012). Under the circumstance, the decision support system, which incorporates the social media data, provides econometric models and supports the decision discussion and presentation, becomes vitally important for policy decision making.

3 THE MIDAS MODELS

The econometric modelling approaches presume the past information can be used to forecast the future values of the variables. According to the data frequency of the explanatory variables and explained variable, they are classified into two kinds. If the variables are sampled at the same frequency, like ARIMA and ADL models (Box and Jenkins 1968; Engle and Granger 1987), belong to the first kind. The models dealing with the variables with different frequencies can be called as mixed data sampling (MIDAS) approach (Ghysels et al. 2005). This part will illustrate three kinds of models: ARIMA, ADL, and MIDAS models. The first two belong to the first kind, and MIDAS is the second type.

An ARIMA( p, d, q ) model is given by: \( (1-\sum_{i=1}^{p} \phi_i L)(1-L)^d Y_t = (1+\sum_{i=1}^{q} \theta_i L)\varepsilon_t \)

Where \( L \) is the lag operator, \( \phi_i \) are the parameters of the autoregressive part of the model, \( \theta_i \) are the parameters of moving average part and \( \varepsilon_t \) are error terms. \( p \) and \( q \) are chosen by AIC or BIC criteria. \( d \) is the differencing order to make the series stationary.

An ADL model is defined as:

\( Y_t = \alpha + \sum_{i=0}^{l} \beta_i X_{t-i} + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t \)

Where \( l \) and \( p \) are the lags of explanatory and explained variables. \( \beta_i, \phi_i \) are the parameters of variables, and \( \varepsilon_t \) are error terms. ADL model contains more information on exogenous variables than AR models.

The MIDAS approach can directly model the mixed frequency data, with a simple, parsimonious and flexible way. The original work on MIDAS focused on volatility predictions. Recently, the method was used to improve quarterly or monthly macroeconomic forecasts with monthly or daily data (Ghysels et al. 2006; Armesto et al. 2010). Ghysels et al. (2007) demonstrated that the standard regression model imposing the equal weights to the variables, yielded asymptotically inefficient and inconsistent estimates, and caused adverse effects on forecasting.

MIDAS regression models are closely related to distributed lag models. The MIDAS approach regresses \( y \) on a distributed lag of \( x \), where \( x \) is sampled at a higher frequency. The MIDAS with lags of the variables, known as ADL-MIDAS (\( p_y, q_x \)) is given by:

\( Y_{t+1}^y = \mu + \sum_{j=0}^{p_y-1} \mu_j Y_{t-j}^y + \beta \sum_{j=0}^{q_x-1} \sum_{i=0}^{N_x-1} w_i^{N_x, j} (\theta^x) X_{t-j}^x + u_{t+1} \)
Where, \( Q \) and \( D \) stand for the data frequency, quarter and daily. \( p^Q_t \), \( q^D_t \) mean the lags of \( Y^Q_t \) and \( X^D_{N_D - j - 1} \cdot \sum_{j=0}^{N_D-1} w_{N_D-j} (\theta^D) \) represents the AR terms of the explained variable. \( N_D \) means the days in a quarter. \( Y^Q_p \), \( D^Q_p \), \( X^D_q \), \( D^D_N \) are the weighing functions, and \( \sum_{j=0}^{N_D-1} w_{N_D-j} (\theta^D) = 1 \). Almon or Beta functions are usually chosen as weighing functions (Ghysels et al. 2006).

Based on the ADL-MIDAS \( (p^Q_t, q^D_t) \), a more complicated MIDAS model, noted as ADL-MIDAS \( (p^Q_t, q^D_t, J^D_X) \), was proposed (Giannone et al. 2008). By adding \( J^D_X \), the model realize the nowcast. The information between \( t \) and \( t+1 \) will be incorporated into the models to forecast the value of \( t+1 \).

The model with \( J^D_X \) leads is noted as:

\[
Y^Q_{t+1} = \mu + \sum_{j=0}^{p^Q-1} \mu_{j+1} Y^Q_{t-j} + \beta \sum_{j=0}^{p^Q-1} \sum_{i=0}^{q^D-1} w_{j+q^D-i+N_D} (\theta^D) X^D_{N_D-j-i-1} + \beta \sum_{j=0}^{p^D-1} w_{j+1} (\theta^D) X^D_{J^D_X-j+i+1} + \epsilon_{t+1}
\]

Where \( J^D_X \) is the lead of the explanatory variables.

MIDAS approach modelling time series with different frequency, which could contain as much useful information of the high frequency part of social media data as possible. It can combine the social media data and the traditional statistical data. In the empirical study, ARIMA, ADL, ADL-MIDAS \( (q^D_t, J^D_X) \) and ADL-MIDAS \( (J^D_X, q^D_t, J^D_X) \) are all used to forecast economic indicators, and evaluated by the criteria.

4 MULTISOURCE MACROECONOMIC FORECAST SYSTEMS DESGIN

4.1 The System Framework

To incorporate the social media data into the macroeconomic forecast system, a framework is proposed. In this framework, multisource data is utilized to forecast the important macroeconomic indicators, combining the social media data with the statistical data. The social media data, which is crawled from different web platforms, provides a comprehensive aggregation of public concerns. These social media data represents the collective understandings and expectations of current and future economic situations. It may influence the message readers’ economic behaviour (such as investment, savings and consumptions), and then further affect the macroeconomic trends in future. Therefore, aggregated opinions on economic situation extracted from social media platforms are timely reflective and predictive indicators of macroeconomic dynamics, which may hopefully improve traditional macroeconomic forecast methods. Fig.1 provides an illustration of the system framework. The system is composed of five modules: knowledge extraction, multisource data management, pre-processing, modelling and forecasting and forecast tracking and evaluation.
(1) Knowledge Extraction

It is necessary to build a knowledge base to extract related news from websites. The knowledge base contains a set of subject-related keywords obtained from experienced experts, authoritative reports etc. which cover economic growth, inflation level and so on. Knowledge base provides support for access to the multisource data by using some techniques like natural language processing.

(2) Multisource Data Management

‘Multi-Source’ is the significant feature of the System, which means not only the official data published by National Bureau of Statistics (NBS) or People’s Bank of China (PBC), but also the social media data from different platforms are incorporated into the system. Multisource data contains four parts. (a) Related news extracted from some important financial websites via the web crawler based on pre-defined keywords from the knowledge base. (b) Reviews and comments extracted from other web platforms, such as blogs, micro-blog etc. (c) Search queries acquired from main search engines, like Google, Baidu and so on. (d) Statistical official data related with the macroeconomic aspects.

(3) Pre-processing

Pre-processing step is necessary before modelling and forecasting. For the texts, it is to extract and acquire the useful information, like time, institution, specific person, main events and the affects and so on. For the time series data, the pre-processing step is to analyze the series outliers, the trends, the seasonality and other significant features. Only when the social media data is strongly correlated with the specific macroeconomic indicator can it be used in the macroeconomic forecasting system.

(4) Modelling and Forecasting

Many modelling techniques are included in the framework, including AR, ADL, and ECM etc. MIDAS model is introduced as a proper technique modelling the mixed frequency data. The high-frequency social media data can be incorporated into the macroeconomic forecast with the help of MIDAS approach. Machine learning techniques such as text-mining, neural networks and support vector machines etc. can also be used as complementary instruments.

(5) Forecast Tracking and Evaluation

It is essential to track the forecast values and evaluate the prediction accuracy. The common methods to track the forecast models are tracking signal and predication-realization diagram. Practically, the
A tracking signal is an indicator that is used when the validity of the forecasting might be in doubt. The prediction-realization diagram is used to compare forecasts with the actual changes that were realized. Besides, the common criteria for evaluating the forecast accuracy include Mean Error (ME), Mean Absolute Error (MAE), Mean Proportional Error (MPE), Mean Absolute Proportional Error (MAPE), and Mean Squared Error (MSE).

4.2 The MIDAS Approach for Economic Forecast

The multisource macroeconomic forecast system combines social media data and the officially statistical data. MIDAS combines the multisource data to forecast the important macroeconomic indicators, which satisfy the macroeconomic forecasting task. As the Fig. 2 shows, a set of mixed frequency time series are used to forecast the important macroeconomic indicators through MIDAS. The multisource social media series are exogenous variables in the modelling, which is updated timely. Therefore, it is not necessary to extrapolate the values for the social media series. However, statistical series, as other exogenous variables, are needed to be extrapolated for forecasting the future values.

![Figure 2. The exogenous variables used to forecast by MIDAS approach](image)

A normal form and some examples instructing how to incorporate the social media data into the macroeconomic forecast are showed in the following.

$$macro_t^L = \alpha + \sum_{i=0}^{h_1} \eta_i macro_{i-1}^L + \sum_{i=0}^{h_2} \beta_i ex_{i-1}^L + \cdots + \sum_{i=0}^{h_n} \beta_{ni} ex_{ni(i-1)}^L + \gamma_i social_{i}^H$$

In the normal form, the left variable is the low-frequency macroeconomic variable, such as $gdp_t^0$, $cpi_t^m$, the letter $L$ means the data frequency, quarter or month. The right variables, $ex_{i-1}^L$, $\cdots$, $ex_{ni(i-1)}^L$ are the low frequency explanatory variables. $h_1, \cdots, h_n$ are the lags of the explanatory variables. $macro_{i-1}^L$ is the $h_i$ lags of the explained variables. $social_{i}^H$ is the high frequency explanatory variable that is extracted from multi data source. Here are some examples on modelling the important macroeconomic indicators.

The GDP forecast model:

$$gdp_t^0 = \alpha + \beta_1 cons_t^0 + \beta_2 asset_t^0 + \beta_3 popu_t^0 + \gamma_1 social_{-cons}^0 + \gamma_2 social_{-gdp}^W$$

The quarterly GDP is explained by the variables based on the economic theories. $cons_t^0$ is the total consumption; $asset_t^0$ represents the fixed assets investment; and $popu_t^0$ is the urban population. $social_{-cons}^0$, $social_{-gdp}^W$ are the social media data related with the consumption and GDP. The
social media data can be the search volumes of the public, the concerns on the economic growth in China, the expectation on the future GDP.

The trade forecast models:

$$ exp_i^M = \alpha + \beta_1 \text{countgdp}_{1i}^M + \beta_2 \text{countgdp}_{2i}^M + \beta_3 \text{exchange}_{3i}^M + \gamma_i \text{social}_\text{exp}_{4i}^W $$

$$ imp_i^M = \alpha + \beta_1 \text{priceratio}_{1i}^M + \beta_2 \text{countgdp}_{1i}^M + \beta_3 \text{countgdp}_{2i}^M + \beta_4 \text{exchange}_{3i}^M + \gamma_i \text{social}_\text{imp}_{4i}^W $$

The models above explain the total import value and total export value based on trade theories. The GDP growth of the major trading partners and the exchange rate influence the export value to a great extent. Besides, the ratio of import price index to the fixed asset investment price index, the economic growth of other countries and exchange rate affect the import value. On that basis, relevant social media data are incorporated to forecast the export and import values.

5 AN EMPIRICAL STUDY ON CPI FORECAST

5.1 Price Index extracted from Google Trends

This part will elaborate how social media data related to macroeconomic aspects are mined, analyzed and use to forecast the future trends of China’s CPI, which is one of the targets for China’s monetary policy. Since 1990, China has arrived at the highest inflation rate in 1994. Especially, CPI reaches 27.7% in October 1994. Considering the high inflation rate, China adopted sound monetary policy to preserve price stability. The price stability is not only the urgent issue for the government, but also draws public attention. The statistical CPI data is published by National Bureau of Statistics monthly and quarterly, usually with a lag of 10 to 15 days respectively.

The social media data related to prices are obtained by Google Trends which is a service provided by Google Inc. It normalizes each time series by dividing the count for each query by the total number of online search queries submitted during the week. The time series data from Google are available at weekly frequency from January 2004 that reflects the concerns on the inflation of China from the web users. The public concerns with prices related with their daily life reflecting the common people’s expectation to the consumer price index in the forthcoming month. The more search volumes of the specific issues is from the search engines, the higher expectation to the current CPI is from the public opinions. The data representing the consumer price in the point of public review can be available immediately at the end of this month. However, the actual statistical CPI data are published in about 15 days of that month. Therefore, the search query volumes from Google Trends are on the behalf of the consumer price expectation that could be used to forecast the statistical CPI data.

The Google search query volumes are weekly data, while the statistical CPI is monthly data. In order to fully use the Google Trends data, the weekly search series data are directly used to the model with the statistical monthly CPI. Given the keyword ‘prices’ (‘wujia’, in Chinese), the Google Trends returns the normalized volume of search queries. The search series is noted as ‘GooPrice’. As there is noise in the series of GooPrice, preprocessing is necessary. GooPrice is presented below after removing the nosing points. Fig.3 displays the social media data related with the CPI from the first week of 2004 to the last week of 2012.
Figure 3. The social media data extracted from Google Trends

In the second step, Augmented Dickey-Fuller test statistic (ADF) method is adopted to test the stationarity of GooPrice. The test result listed below shows that GooPrice is a stationary series.

<table>
<thead>
<tr>
<th>Augmented Dickey-Fuller test statistic</th>
<th>t-Statistic</th>
<th>Prob.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Augmented Dickey-Fuller test statistic</td>
<td>-4.5294</td>
<td>0.0002</td>
</tr>
<tr>
<td>Test critical values:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1% level</td>
<td>-3.4444</td>
<td></td>
</tr>
<tr>
<td>5% level</td>
<td>-2.8676</td>
<td></td>
</tr>
<tr>
<td>10% level</td>
<td>-2.5701</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. The stationarity test of GooPrice series

To analyze the two series clearly, the aggregation method for converting the weekly GooPrice into the monthly GooPriceM is necessary. There is seasonality in the GooPriceM, so the third step is to extract the cycle item of GooPriceM. X12ARIMA is used to eliminate the seasonality and the noise, then, HP filter is adopted to extract the cycle item. Fig.5 shows the two time series.

Figure 5. The monthly Google search series and the statistical CPI

GooPriceMCycle is strongly correlated with the CPI series with the correlation coefficient is 0.793. There is something interesting with the GooPriceM Cycle before and after the financial crisis burst in 2007-2008, before the financial crisis, the two series are almost synchronized. However, after the crisis, the GooPriceM Cycle leads to the CPI by above six months. One potential explanation is that after the crisis, the public are more concerned with the changes of prices via the Internet retrieval. Consequently, the expectation on the future price contributes a lot to the real price trends. Thereafter, the GooPriceM Cycle reflecting the public concerns leads to the CPI price since 2008.
After all the preprocessing steps, GooPrice and GooPriceMCycle, both weekly and monthly series are available, which will be used to forecast the monthly CPI in next section.

5.2 The MIDAS Model for CPI Forecast

Since the CPI series is not stable, the difference is employed to make the data series stable. The existing researches have taken ARIMA model as the benchmark model in forecasting CPI. ARIMA model is very simple, classical, and easy to conduct. It uses the past information of the time series to forecast the future values. Therefore, many literatures use it as the benchmark model to examine the effectiveness of their own complicated approach.

Here, three models are constructed to forecast the statistical CPI data series. The sample is from 2005.1 to 2012.11. The left side variable is the monthly CPI, and the right side variables are the weekly GooPrice, and monthly GooPriceMCycle.

Model 1:
\[
d_{t}^M = 0.27d_{t-1}^M + 0.24d_{t-3}^M - 0.29d_{t-12}^M - 0.91\epsilon_{t-12}^M \\
0.0203 0.0421 0.0055 0.0000
\]
\[
y_{t}^M = y_{t-1}^M + d_{t}^M
\]

Model 2:
\[
y_{t}^M = 3.45 + 0.93y_{t-1}^M - 0.91\epsilon_{t-12}^M + 0.55GooPriceMCycle_{t-2}^M \\
0.0000 0.0000 0.0000 0.0003(p-value)
\]

Model 3:
(M3.1) model with the whole month: ADL-MIDAS(1,’1m’)
(M3.2) model with a lead of 2 weeks: ADL-MIDAS(1,’1m’,’1w’)
(M3.3) model with a lead of 3 weeks: ADL-MIDAS(1,’1m’,’3w’)
(M3.4) model with a lead of 1 week: ADL-MIDAS(1,’1m’,’1w’)

Model 1 is selected as the benchmark model, and the left side is the differential term of monthly CPI that is used to eliminate the non-stationarity. The right sides are AR and MA items, and the lags are chosen by AIC or BIC criteria. Model 2 selects the integrated monthly series, GooPriceMCycle, which is the average of the whole weeks in one month. The numbers below Model 2 are P-values of the coefficient.

The detailed MIDAS model description is not listed here because of the complex forms. Model 3 contains four sub-models, and each model represents the new information used for forecasting the future value of CPI. M3.1 means that the whole weekly information can be employed to forecast the future values of CPI. And M3.2 to M3.4 shows that 1 week, 2 weeks and 3 weeks are used to forecast the whole monthly value. The forecasting sample is from 2012.1 to 2012.11. The figure below shows part of the results.
Table 7. The precision criteria of the forecasting models

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3.1</th>
<th>Model 3.2</th>
<th>Model 3.3</th>
<th>Model 3.4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>0.1640</td>
<td>-0.1502</td>
<td><strong>0.0265</strong></td>
<td>-0.1444</td>
<td>0.1110</td>
<td>-0.2130</td>
</tr>
<tr>
<td>MAE</td>
<td>0.7060</td>
<td>0.8147</td>
<td>0.0818</td>
<td>0.2561</td>
<td><strong>0.0271</strong></td>
<td>0.3567</td>
</tr>
<tr>
<td>MPE</td>
<td>0.0879</td>
<td>-0.0271</td>
<td><strong>0.0106</strong></td>
<td>-0.0547</td>
<td>0.0506</td>
<td>-0.0837</td>
</tr>
<tr>
<td>MAPE</td>
<td>0.1541</td>
<td>0.1268</td>
<td><strong>0.0108</strong></td>
<td>0.0573</td>
<td>0.0506</td>
<td>0.0862</td>
</tr>
<tr>
<td>MSE</td>
<td>0.2576</td>
<td>0.1881</td>
<td><strong>0.0012</strong></td>
<td>0.0265</td>
<td>0.0140</td>
<td>0.0533</td>
</tr>
</tbody>
</table>

All the criteria evaluate the forecast accuracy rate from different views. The criterion MSE reports that MIDAS models, Model 3.1 to Model 3.4, outperform the Model 1 and Model 2 completely. Model 3.1 is the best one according to this criterion. That is because the Model 3.1 contains the whole information of one month, which is useful to forecast. The criterion of MAE shows that Model 3.3 outperforms other models. It indicates that the model incorporating the information of 2 weeks can also improve the forecast accuracy. Other criteria demonstrate that sets of Model 3 are better than the benchmark model and ADL model. The Model 3.1 reduces the ME, MAE, MPE, MAPE, and MSE by average 90% compared to the benchmark model. In a word, it is proved that incorporating the weekly social media data related with price could help forecast the real statistical CPI.

### 6 CONCLUSION

This research introduces a mixed data sampling approach to incorporate the high frequency part of the social media data into the econometric models, and proposes a decision support system framework for multisource macroeconomic forecast. To validate the proposed framework, an empirical study on China’s CPI forecast is then conducted. In the empirical study, the social media data related to prices is extracted from Google Trends, pre-processed, and used to construct the forecast models. The empirical results indicate that the MIDAS model incorporating the weekly social media data outperforms the benchmark model and the ADL model. In addition, the MIDAS model using the information obtained from the first two weeks of one month can already outperform the models without incorporating the social media data.

This paper attempts to reveal the huge potential of social media data in the macroeconomic forecast, and offer a feasible solution on the application of social media data in the big data era. Since the social media data is updated timely, it can predict the future values of the economic indicators ahead of the release of statistical data. Moreover, the directly incorporation of the social media data could improve the forecasting accuracy to a great extent by offering more tangible clues of what people perceive about the economic situation. Therefore, the social media data makes the macroeconomic forecast
more timely and accurate, which will be extremely helpful for the policy decision makers and enterprises. The work contributes to the recent literatures on dealing with the social media data by introducing a mixed modelling approach, which could make the best use of the high frequency part of social media data. Results of this research will have great implications on future development in social media data extraction and modelling approaches for macroeconomic forecasting in the big data era.

However, this research used Google trends data as a convenient way for capturing the trends of social media opinions in China’s CPI forecasting. More social media data sources and more forecasting targets would help to further validate and advance the performance and generalizability of the proposed mixed frequency modelling approach. In future researches, more economic indices need to be extracted from the social media data to support the comprehensive macroeconomic forecast tasks. In the data pre-processing methods, more text-mining and semantic-based modelling techniques are needed to be employed for the multisource social media data processing. Considering the subjectivity and complexity of the social media data, the sentiment analysis and the network analysis techniques will also be incorporated into the multisource macroeconomic forecast framework. In the empirical studies, other macroeconomic issues will be researched, including the economic growth, trade contracts and investment strategy and so on. We hope this research not only promote the inclusion of social media data in traditional macroeconomic forecasting based on econometric modelling, but also contribute to the social media data research communities to adapt economics theories based methodologies to the massive data analysis to support business decisions.

Acknowledgements

This research is sponsored by National Science Foundation of China (Project No.: 71171186, 91224006). The authors would like to thank the two anonymous referees for their helpful comments and suggestions.
References


