A macroeconomic monitoring index construction management method based on big data analysis

Liangyoutong Li*

Affiliated School of Kunming Nsau Research Institute, Kunming, Yunnan, China

*Corresponding author: 2103084897@qq.com

Abstract. In order to effectively forecast and monitor China's price data, the article takes CPI and PPI price series as an example, constructs CPI and PPI high-frequency monitoring indices using commodity price big data, collects high-frequency big commodity big data from January 1, 2009 to December 27, 2019, and constructs a mixed-frequency sampling model (MIDAS). The results show that the mixed-frequency MIDAS model with big data has better dynamic forecasting effect on CPI and PPI than the traditional ADL and GARCH models, which proves that the mixed-frequency MIDAS model with big data has better monitoring effect on China's price data.

Keywords: CPI; PPI; big data; MIDAS; mixed-frequency data

1 Introduction

The main objectives of macroeconomic policy-making are to promote economic growth, ensure full employment and maintain price stability. In order to achieve these goals, governments, companies, or relevant departments need to monitor and summarize the current macroeconomic performance in a timely manner, and at the same time forecast the future short-term trend, so as to adopt appropriate regulatory policies to cope with possible economic recessions and economic turmoil [1].

However, the current high cost of survey data has resulted in most of the official data published by the Bureau of Statistics being monthly or quarterly reports, such as monthly PMI, consumption, total investment, total import and export value, quarterly GDP, etc. are low-frequency data that cannot be used to observe and track short-term changes in economic performance, and if low-frequency data are used to analyze changes in financial market performance, they cannot reflect the latest situation [2-3]. Low-frequency data ignore too many details and mostly focus on downstream indicators, which reflect more the results of economic and financial operations and cannot observe the process, which is not conducive to the comprehensive monitoring of the economic system [4]. At present, economic data of each country are still mainly obtained through traditional survey methods, with the defects of excessive human error, certain time lag, huge cost, low reliability, and mostly unidimensional data.

X. Ding et al. (eds.), *Proceedings of the 2023 4th International Conference on Big Data and Social Sciences (ICBDSS 2023)*, Atlantis Highlights in Social Sciences, Education and Humanities 12,
https://doi.org/10.2991/978-94-6463-276-7_8

High-frequency data usually have a frequency of less than daily, presenting a smoother historical curve and facilitating the observation of short-term changes in financial markets. Since high-frequency data itself is not an independent indicator system and is dependent on low-frequency data indicators, we need to properly handle the relationship between low-frequency data and high-frequency data [5]. Secondly, due to the higher frequency of high-frequency data, its autocorrelation will be stronger, which in turn makes high-frequency indicators present the characteristics of volatility aggregation.

A new generation of data collection and processing technology - big Data technology [6]. This technology can help researchers collect massive, high-frequency and multidimensional data [7]. These data can make up for the shortcomings of existing traditional statistical methods and help researchers better judge the economic situation. The introduction of big data can effectively make up for the shortcomings of traditional survey methods, and big data methods have the characteristics of wide range of data collection, low cost, real-time reflection of economic conditions, and low human error. A number of foreign scholars and relevant Many foreign scholars and relevant government departments have built a reliable real-time economic data system through the network big data method, but there is a lack of relevant research results in China [8]. In order to make up for the shortcomings of China's statistical system in terms of real-time and reliability, this paper plans to use web crawler technology to collect web big data, and clean and classify the web data to build a reliable real-time data system of China's private sector economic activities [9].

In order to effectively use high-frequency big data for accurate forecasting and high-frequency monitoring, more advanced econometric models are needed. In recent years, measurement models based on mixed-frequency data have received more and more attention, the MIDAS (mixed-data sampling) model and demonstrated the effectiveness of this method for a variety of different forecasting objectives. Unlike traditional prediction models, MIDAS can be considered as a regression model that allows the explanatory variables to be high-frequency data and the explanatory variables to be low-frequency data. The model can effectively solve the problem of low-frequency data not being able to track short-term economic and financial changes, thus acting out the high-frequency data and downscaling the high-frequency explanatory variables using a distributed lag polynomial weight function to control the number of coefficients to be estimated. Initially MIDAS was used for volatility forecasting, while later the method was used for forecasting macroeconomic series and used monthly indicators for quarterly GDP forecasting [10-11].

The purpose of this paper is to construct a MIDAS model with big data on commodity prices to forecast and monitor CPI and PPI data in China. It is important to use big data to forecast and monitor macroeconomic data.

2 Model Construction

This paper constructs a CPI forecasting model for China using high-frequency big data based on mixed-data sampling (MIDAS).

The traditional forecasting method uses the same frequency variables as the predicted object. The traditional forecasting method uses the same frequency variables as the forecasted object. The ADL (autoregressive distribution lag) model is used as an example:

$$Y_t = \mathrm{a} + \beta(\mathrm{L})Y_t + \gamma(\mathrm{L})X_t + \quad _t \tag{1}$$

where Y_t is the predicted variable, X_t is the predictor variable with the same frequency as Y_t used to make the prediction. β (L) and γ(L) denote the weighted average of the variable lag the weighted average of the terms,

$$\beta(L) = \sum_{i=1}^{p} b_i L^i,\ \gamma(L) = \sum_{i=1}^{q} g_i L^i \tag{2}$$

In brief, the traditional ADL forecasting method selects the same frequency data and uses the weighted average of the lagged variables for forecasting.

The problem with traditional forecasting models is that the frequency of economic indicators is inconsistent, so this paper emphasizes the use of high-frequency mixed-frequency forecasting models in macroeconomic testing. In the case of CPI forecasting, for example, the current month's CPI data is usually released in the middle of the following month, and the timeliness of the data is likely to be lost during the data collection process. If traditional forecasting methods are used, the value of the monthly predictor variable (X_t) needs to be known. However, the predictor variables are usually also released in the middle of the following month, which poses a significant challenge to the timeliness of the forecast. Moreover, if we use only same-frequency data for forecasting, a lot of valid information will be missed, so we must integrate and use the model to convert to high-frequency data.

In order to make full use of all available data in forecasting, it is necessary to use big data, which is high-frequency data, and to use high-frequency data to forecast low-frequency data, it is necessary to introduce the mixed-frequency MIDAS model.

The prediction idea of MIDAS is to use all available data at time tm for prediction. The general expression is:

$$Y_{t_m+h_m} = \beta_0 + \beta_q(L)Y_{t_q} + \sum_{i}^{N} \gamma_{i,m}(L,\theta)X_{i,t_m} + \varepsilon_{t_m} + h_m \tag{3}$$

Among them, $X_{i\,,\,t_m}$ represents the value of the i-th predictor variable in period t_m. In the estimation model, the data units are first standardized to the highest frequency unit because of the mixed frequency data. Since the highest frequency of the data involved in this model is day, the data release time is normalized to day. For example, assuming that the January 2019 predicted data is released on February 10, it is expressed as the 40th day of 2019 data. Suppose that on day t_m, need to firecast the data $Y_{tm}+h_m$ published on day $t_m + h_m$, with hm denoting the forecast horizon. Then, in order to make full use of the available data for the forecast, we will use:

(1) The data at day t_m and all availabl Y_t data for the lag period. Y_{tq} denotes the most recent period of published data to date t_m . Using the lagged polynomial parameter estimation using the lagged polynomial multiplication method, and after obtaining the parameter estimates, the model prediction mean β_q (L) = $\sum_{i=1}^{p} b_i L_q^i$ is weighted to the lagged values of the data.

(2) Data at day t_m and all available X_t data for the lagged period. In order to solve the problem of covariance and low degrees of freedom caused by too many parameters to be estimated, this paper the Almon polynomial γ_m (L,θ) = $\sum_{i=0}^{q} c(i, \theta) L_m^i$ is used to weight the lag series of Xt_m of the lag series is weighted. Where

$$c(i,\theta) = \exp(\theta_1 i + \theta_2 i^2 + \ldots + \theta_q i^q) / \sum_{i=0}^{q} \exp(\theta_1 i + \theta_2 i^2 + \ldots + \theta_q i^q) \quad (4)$$

Multiplication method for parameter estimation, and after obtaining the parameter estimates, the model prediction mean value $E[Y_{tm}+h_m|t_m] = \beta_0 + \beta q\ (L)Y_{tq} + \gamma_m\ (L, \theta\ )X_{tm}$. As at the moment tm for the span hm variables after Ytm +hm It's real-time forecasting $E[Y_{tm} +h_m\ |t_m]$ is a higher-frequency index that is highly consistent with official data and can be used as a high-frequency monitoring index. Of course, the monitoring effectiveness of the index depends on the forecasting effect. If the forecasting effect is better and the data is more accurate, then the index's monitoring of the economic situation will be more detailed and accurate, and the effect will be better.

3 Result Analysis

3.1 CPI forecast results

This paper first uses the MIDAS mixed-frequency model to predict the monthly CPI growth rate using mixed-frequency big data. The mixed-frequency data collected and used in predicting CPI include: (1) Daily risk-free interest rate data (represented by the weighted average interest rate of interbank lending within 7 days) that characterizes the degree of loose liquidity in the financial market, M2 growth lagged in January speed. (2) China's Ministry of Commerce's China Commodity Weekly Index (including general index, iron and steel, energy, non-ferrous metals, minerals, oils and fats, rubber, and sugar) used to describe energy transportation prices, a total of 8 class data). (3) Keqiao textile price index used to describe clothing prices (including prices of raw materials such as cotton, linen, viscose, and polyester; prices of gray fabrics such as natural fibers, chemical fibers, and blended fibers; prices of clothing fabrics such as pure cotton and pure linen; Prices of home textiles such as bedding and daily-use home textiles, prices of clothing accessories such as thread and lace, a total of 31 categories of prices). (4) The weekly price data of edible agricultural products released by the Ministry of Commerce of China (including chicken, duck, geese, poultry, edible oil, eggs, rice, flour and other grains, soybean oil, peanut oil, rapeseed oil and other oils, pork and mutton , beef and other meat, cucumber, cabbage, white radish and other vegetables,

carp, grass carp, yellow croaker and other aquatic products, apples, bananas, watermelons, ducks, pears and other fruits, a total of 43 kinds of agricultural product data), released by the Ministry of Commerce The price data of edible agricultural products basically include the types of agricultural products in the CPI basket of commodities. (5) Jingdong network e-commerce data. (6) Dummy variables (used to describe the Spring Festival and holiday factors). The data range is from January 1, 2009 to November 27, 2019.

In this paper, the JD.com e-commerce daily data is added to the model. For JD.com's online e-commerce data, firstly, according to the categories of a basket of commodities in China's CPI statistics, the products on JD.com's platform are classified, so that the online retail price index constructed in this paper is comparable to the CPI index. The China CPI survey catalog is determined according to the national urban and rural areas and residents' consumption structure and consumption habits, and is divided into 8 categories and 262 basic categories according to purposes. From 2000 to 2015, the survey content was classified into food, tobacco, alcohol and supplies, clothing, household equipment supplies and maintenance services, medical care and personal supplies, transportation and communications, entertainment, education and cultural supplies and services, and residence; starting in January 2016 Using 2015 as the new round of comparison base period (rotated every five years), the CPI survey catalog has also been adjusted, mainly including: (1) The original "food" and "tobacco and alcohol" are merged into the current "food, tobacco and alcohol" ; (2) The original "health care and personal products" was split into the current "daily goods and services", "health care" and "other products and services"; (3) the original "entertainment, education, cultural products and services" ” was split into what is now “educational culture and entertainment” and “other supplies and services”; (4) The original "household equipment and maintenance services" was split into the current "daily goods and services" and "other supplies and services"; (5) "food" in the old classification was a major category, including food , Meat and Poultry, Fresh Vegetables, Fresh Fruits, Aquatic Products, Tea and Beverages, Dining Out, etc.; the new “Food” is the middle category under the category of “Food, Tobacco and Alcohol”, which only includes grain, livestock meat, poultry meat, fresh vegetables, and fresh fruits , aquatic products, etc., no longer include "tea and beverages" and "catering outside the home"; (6) newly added "horticultural flowers and supplies", "pets and supplies", "elderly care services" and "financial services" and other household expenditures Add faster categories.

In view of the above considerations, this paper divides online retail commodities into 19 categories, including rice noodles, edible oil, vegetables, chicken, pork, mutton, beef, duck, seafood, milk and dairy beverages, fruits, liquor, beer, wine, Men's shoes, men's clothing, women's shoes, women's clothing, children's clothing.

This paper uses web crawler technology to directly collect high-dimensional data such as sales prices and sales of commodities from the Jingdong online retail platform. First construct a price index for each category, the index construction method is as follows:

$$\mathrm{pi} = \sum_{j=1}^{n_i} p_{ij} \frac{q_{ij}}{\sum_{j=1}^{n_i} q_{ij}} \tag{5}$$

P_i is the price index of the ith commodity category constructed in this paper p_{ij} and q_{ij} are the price and sales volume of the jth commodity under this category, respectively, and n_i represents the total number of commodities in the ith commodity category, under the collection methodology of this paper, n_i Generally around 10000 kinds of goods. The price index of this category is obtained by weighting the products with different prices of all commodities under the same category according to their sales volume. The advantage of this method is that the consumption habits of consumers are fully considered in the weighting of commodities, and the problem of missing commodities in the traditional CPI index construction is effectively avoided during the sampling process, which improves the accuracy of the data.

The dynamic prediction effect of the CPI-Midas model in this paper is shown in Figure (with $h_m = 7$ 1the prediction span is one week).

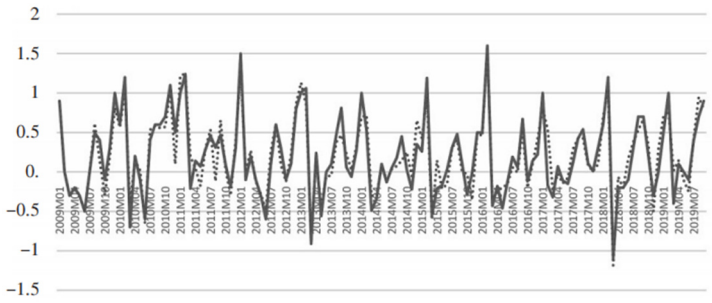

Fig. 1. Dynamic prediction effect of big data MIDAS model on CPI ($h_m = 7$)

As can be seen from Figure 1, under dynamic forecasting, the big data MIDAS model has a very good forecasting effect, basically predicting each inflection point of the CPI. The mean absolute error (Mean Absolute Error, MAE) of the dynamic prediction of the model is used to describe the prediction effect of the model. In this paper, we set different prediction spans $h_m = 7$, 14 , 21, the prediction spans are 1,2 and 3weeks， and compare the prediction effect of the big data MI- DAS model with the traditional ADL and Garch models.

Table 1. Predictive effects of the models (MAE)

model	$h_m = 7$	$h_m = 14$	$h_m = 21$
ADL	0.381	0.381	0.381
GARCH	0.392	0.392	0.392
ARIMA	0.412	0.412	0.412
VAR	0.412	0.412	0.412
Midas	0.127	0.132	0.141

Since the traditional ADL cannot utilize high-frequency data, the results in Table 1 show that the dynamic prediction errors of ADL, GARCH, ARIMA, and VAR are the same under different prediction spans. By observing the data, it can be found that the prediction effect of the MIDAS model using high-frequency big data is better than that of the traditional GARCH, ADL, ARIMA and VAR models, and the conclusions are valid under different prediction spans, and the conclusions are robust, this demonstrates the feasibility of the idea in this paper to use the predicted value of the MIDAS model, $E[Y_{tm} + h_m | t_m]$as a high-frequency CPI monitoring index.

3.2 PPI forecast results

PPI represents the price index of industrial producers, which is different from the consumer price index represented by CPI: (1) The weights of means of production and means of living in PPI are 0.75 and 0.25, and more than 95% of the changes in PPI are caused by means of production Therefore, for the prediction of PPI, the data used by the MIDAS model is biased towards industrial products; (2) If classified according to industrial industries, the PPI price index includes 39 major industries, covering mining, manufacturing, electricity, gas and water There are 3 industrial categories of production and supply industry. Based on the above observations, the data used in the mixed-frequency data model in this paper include: (1) For the research on the price of living materials, the model selects the average price data series of daily industrial consumer goods in 36 large and medium-sized cities, including about 30 commodity prices ;(2) The price index series of agricultural means of production, including the prices of 10 kinds of agricultural means of production; (3) China's Ministry of Commerce's weekly commodity price index (general index, steel, energy, non-ferrous metals) used to describe energy prices Metals, minerals, oils and fats, rubber, sugar, a total of 8 types of data); (4) Nanhua Futures Commodity Daily Index, including futures prices of about 40 industrial products and agricultural products. The data range is from January 1, 2009 to November 27, 2019.

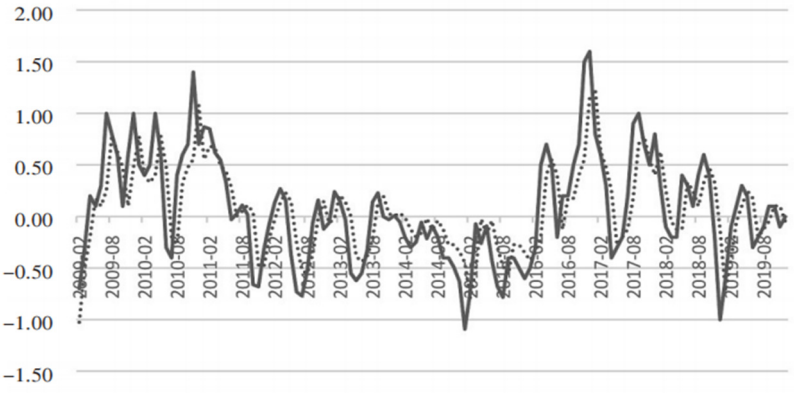

Fig. 2. Dynamic prediction effect of big data MIDAS model onPPI (h_m = 7)

As can be seen from Figure 2, under dynamic forecasting, the big data MIDAS model has a very good forecasting effect, basically predicting the inflection points of the PPI chain growth rate. Similarly, this paper compares the prediction effect of the big data MIDAS model with traditional ADL, Garch and other models.

Table 2. predictive effectiveness of the models (MAE)

models	$h_m = 7$	$h_m = 14$	$h_m = 21$
ADL	0.324	0.324	0.324
GARCH	0.251	0.251	0.251
ARIMA	0.222	0.222	0.222
VAR	0.239	0.239	0.239
Midas	0.124	0.131	0.138

As can be seen from Table 2, the prediction effect of the MIDAS model using high-frequency big data is better than that of the traditional GARCH and ADL models, indicating that the idea of this paper is feasible, and the prediction value of the MIDAS model, $E[Y_{tm} + h_m | t_m]$ can be used as a high-frequency PPI monitoring index.

4 Conclusion

In making macroeconomic decisions, it is necessary to monitor and summarize the current macroeconomic operating situation promptly, as well as to forecast future short-term trends and adopt appropriate regulatory policies to cope with possible economic downturns and turbulence. However, most of the official data released by the Bureau of Statistics are monthly or quarterly data, which are low-frequency data. Low-frequency data ignore too many details and are not conducive to comprehensive monitoring of the economic system. To effectively monitor China's price data, this paper takes the CPI and PPI price series as an example and utilizes commodity price big data to construct CPI and PPI high-frequency monitoring indexes. In this paper, the high-frequency bulk commodity big data from January 1, 2009, to November 27, 2019, are collected, and the mixed-frequency sampling model (MIDAS) is constructed. The dynamic prediction effect of the mixed-frequency MIDAS model for CPI and PPI under big data is better than that of the traditional ADL model and GARCH model, which proves that the mixed-frequency MIDAS model under big data has a better monitoring effect on China's price data.

Reference

1. Brynjolfsson E, Smith M D. Frictionless Commerce? A Comparison of Internet and Conventional Retailers [J]. Management Science,2000,46 (4).
2. Cavallo A. Online and Official Price Indexes:Measuring Argentina's Inflation [J].Journal of Monetary Economics,2013,60(2).
3. Ellison G, Ellison S F. Search,Obfuscation and Price Elasticities on the Internet [J].Econometrica,2009,77(2).

4. Gorodnichenko Y, Sheremirov V, Talavera O. Price Setting in Online Markets: Does IT Click? [R]. NBER Working Paper Series,2014.
5. Nygaard R. The Use of Online Prices in the Norwegian Consumer Price Index [J]. Statistics Norway,2015.
6. Ghysels E, Sinko A, Valkanov R. MIDAS Regressions: Further Results and New Directions [J]. Econometric Reviews,2007,26(1).
7. Krsinich F. Price Indexes From Online Data Using the Fixed-effect Window-splice (FEWS) Index [J]. Statistics New Zealand,2015.
8. Valentino-DeVries J, Singer-Vine J, Soltani A. Websites Vary Prices, Deals Based on Users' Information [J]. Wall Street Journal,2012.
9. Yumeng Tian. 2021. Construction of China's Financial Condition Index and Analysis of Market Early Warning. In The 2021 12th International Conference on E-business, Management and Economics (ICEME 2021). Association for Computing Machinery, New York, NY, USA, 210–217. https://doi.org/10.1145/3481127.3481173
10. Aihua Li, Qiyuan Zhan, Weijia Xu, and Yuejin Zhang. 2022. Research on Tourism Prosperity Index Based on the Power Big Data. In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '21). Association for Computing Machinery, New York, NY, USA, 347–352. https://doi.org/10.1145/3498851.3498977
11. Tao Li. 2018. Using Big Data Analytics to Build Prosperity Index of Transportation Market. In Proceedings of the 4th ACM SIGSPATIAL International Workshop on Safety and Resilience (Safety and Resilience'18). Association for Computing Machinery, New York, NY, USA, Article 17, 1–6. https://doi.org/10.1145/3284103.3284123