

Article

The Prediction of Soybean Price in China Based on a Mixed Data Sampling–Support Vector Regression Model

Xing Liu, Wenhuan Zhou, Zhihang Gao, Dongqing Zhang * and Kaiping Ma

College of Information Management, Nanjing Agricultural University, Nanjing 210095, China; 31322109@stu.njau.edu.cn (X.L.); 31322120@stu.njau.edu.cn (W.Z.); 31322228@stu.njau.edu.cn (Z.G.); makaiping@njau.edu.cn (K.M.)

* Correspondence: zhangdq@njau.edu.cn

Abstract: Soybean is a crucial economic crop and it is one of the most marketized and internationalized bulk agricultural products in China. As fluctuations in soybean prices directly impact national food security and agrarian stability, it is essential to predict this price accurately. Soybean price is influenced by multiple factors, such as macroeconomic data (typically low-frequency, measured quarterly or monthly), weather conditions, and investor sentiment data (high-frequency, for example, daily). In order to incorporate mixed-frequency data into a forecasting model, the Mixed Data Sampling (MIDAS) model was employed. Given the complexity and nonlinearity of soybean price fluctuations, machine learning techniques were adopted. Therefore, a MIDAS-SVR model (combining the MIDAS model and support vector regression) is proposed in this paper, which can capture the nonlinear and non-stationary patterns of soybean prices. Data on the soybean price in China (January 2012–January 2024) were analyzed and the mean absolute percentage error (MAPE) of the MIDAS-SVR model was 1.71%, which demonstrates that the MIDAS-SVR model proposed in this paper is effective. However, this study is limited to a single time series, and further validation across diverse datasets is needed to confirm generalizability.

Keywords: soybean prices; mixed-data sampling (MIDAS); support vector regression (SVR); machine learning



Academic Editor: Dongping Zhu

Received: 28 April 2025

Revised: 20 May 2025

Accepted: 22 May 2025

Published: 26 May 2025

Citation: Liu, X.; Zhou, W.; Gao, Z.; Zhang, D.; Ma, K. The Prediction of Soybean Price in China Based on a Mixed Data Sampling–Support Vector Regression Model. *Mathematics* **2025**, *13*, 1759. <https://doi.org/10.3390/math13111759>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

MSC: 62P20

1. Introduction

Recent trends in global food market instability, combined with the ongoing restructuring of China's agricultural production methods, have underscored the necessity of ensuring self-sufficiency in terms of food. This is especially important for key agricultural products like soybeans, which play a vital role in China's food security. The 2024 Central Government Document No. 1 explicitly emphasizes the need to safeguard food security and maintain subsidies for corn and soybean producers in China [1]. These policies aim to incentivize soybean cultivation and expand planting areas, thereby influencing soybean market dynamics and price trends.

China's reliance on imported soybeans further highlights the importance of price stability. Data provided by China Customs indicates that soybean imports in 2024 reached 105 million tons, accounting for approximately 50% of total domestic consumption. This dependence means that price fluctuations directly impact national food security and agricultural stability. Accurate soybean price predictions enable policymakers to formulate

effective strategies, mitigate market risks for farmers, and increase economic benefits, thus contributing significantly to national development.

Various studies have employed different forecasting methods for soybean prices. Zhu et al. [2] developed a soybean price forecasting model using an enhanced GM (1,1) model based on historical data. Xiong et al. [3] proposed a dynamic model averaging (DMA) framework, identifying time-varying factors in soybean futures prices from both market and economic perspectives. Yang et al. [4] proposed an optimized EEMD-SVR integrated forecasting method, while Zhang et al. [5] developed a QR-RBF neural network model featuring indicators such as the consumer price index and soybean import volumes. Fan et al. [6] created an LSTM model using daily closing prices.

However, most existing studies rely on single-frequency data for their forecasting models, whereas, in practice, a number of different, mixed-frequency factors influence soybean prices. These include macroeconomic data, which is typically reported at quarterly and monthly intervals, and such high-frequency data as weather conditions and investor sentiment, which are generally assessed on a daily basis. Traditional prediction models are unable to analyze variables of differing frequencies directly, requiring preprocessing to align them first; this is often carried out using interpolation methods that convert low-frequency data to high-frequency data, or by aggregating high-frequency data into low-frequency data. These processes may result in significant information loss due to the reduction in sample size. To address this, Ghysels et al. [7] proposed a mixed-frequency data sampling (MIDAS) model, which integrates multi-frequency sample data into a unified framework. This improves prediction accuracy and speed by leveraging valuable information from high-frequency variables to improve the interpretation of low-frequency ones [8]. Guo and Ma [9] employed a Markov-switching mixed-frequency model (MS-MIDAS), which outperforms the standard MIDAS model in terms of predictive accuracy. Cai et al. [10] introduced a novel reverse mixed-frequency data sampling model (R-MIDAS), which features a machine learning algorithm, and demonstrated superior performance across a number of different industries.

Despite its broad range of economic applications, the MIDAS model is rarely used in soybean price forecasting. Wang et al. extended the GARCH-MIDAS method, constructing the GARCH-MIDAS-W and GARCH-MIDAS-W-MBB models, introducing soybean volatility forecasting [11]. Yu et al. explored the impact of meteorological disasters on the volatility of soybean futures returns using a mixed-frequency GARCH-MIDAS model and its extensions to capture short- and long-term market volatility [12]. The current study aims to bridge this gap by employing a mixed-frequency data sampling method for the modeling and analysis of soybean prices.

This paper addresses the complexity of soybean price fluctuations by integrating machine learning into the MIDAS model, enabling the exploration of nonlinear, discontinuous information [13,14]. By analyzing historical data, the proposed approach aims to more accurately capture both short-term fluctuations and long-term trends in soybean prices, leveraging the combined advantages of high- and low-frequency information.

2. Materials and Methods

2.1. Mixed Data Sampling (MIDAS) Model

The univariate MIDAS model is the fundamental form of the MIDAS framework, using low-frequency data as the explanatory variable and high-frequency data as the dependent variable. It creates a regression model based on polynomial weights, estimating

the required parameters through a nonlinear least squares method [15]. The fundamental univariate MIDAS expression is outlined as follows:

$$Y_{t+h} = \beta_0 + \beta_1 W\left(L^{\frac{1}{m}}; \theta\right) X_t^{(m)} + \varepsilon_{t+h} \quad (1)$$

where

Y_{t+h} is a low-frequency dependent variable at period $t + h$.

$X_t^{(m)}$ is a high-frequency predictor at period t .

h is the lead time for high-frequency data to predict low-frequency outcomes.

$W\left(L^{\frac{1}{m}}; \theta\right) = \sum_{i=0}^K \omega(i; \theta) L^{\frac{1}{m}}$ is a weighted lag polynomial, where $L^{\frac{1}{m}}$ denotes the lag operator.

Thus, $L^{\frac{1}{m}} X_t^{(m)} = X_{t-\frac{1}{m}}^{(m)}$ and $\omega(i; \theta)$ denotes the weight function for lagged terms.

m is the frequency ratio (high-to-low frequency).

K is the maximum lag order.

ε_{t+h} is the random error $\sim N(0, \sigma^2)$.

Five types of weighting functions are commonly used—the two Beta density functions (which include Beta-MIDAS and Beta Non-Zero-MIDAS), the Almon function, the exponential Almon function, and the segmentation function [16]. The choice of weighting function frequently depends on subjective judgment, with no uniform standard defined; the specific characteristics of the data have a significant influence on its effectiveness. When the data change, the initial weighting function applied may no longer be applicable. In response to the associated issues, researchers have developed a weight-constrained MIDAS model alongside the unrestricted model. The predictive performances of various MIDAS models have also been assessed, suggesting that the unrestricted MIDAS model demonstrates the best performance [17,18]. The unrestricted mixed-frequency sampling model (U-MIDAS) is defined as follows:

$$Y_{t+h} = \beta_0 + \beta_1 X_t^{(m)} + \varepsilon_{t+h} \quad (2)$$

This model imposes no constraints on the weights assigned to high-frequency explanatory variables, employing standard linear regression to estimate parameters and allowing for dynamic adjustments to the model structure based on data characteristics, offering good adaptability.

The relationship between explanatory variables and explained variables often requires more than one explanatory variable to adequately represent the changes in the explained variable. It is therefore of great importance to develop a multivariate unrestricted mixed-frequency sampling model, which can be expressed as follows:

$$Y_{t+h} = g(x) = \beta_0 + \sum_{i=1}^n \beta_i X_{t,i}^{(m_i)} + \varepsilon_{t+h} \quad (3)$$

In Equation (3), $X_{t,i}^{(m_i)}$ is the i -th explanatory variable in period t , while β_i is the coefficient of the i -th explanatory variable.

2.2. Support Vector Regression (SVR) Model

The SVR method maps the original input data to a high-dimensional space, transforming the original nonlinear problems into linear ones via kernel functions, thereby defining an optimal linear hyperplane. This means that the challenge of identifying the optimal hyperplane is transformed into the actual optimal solution. The introduction of an insensitive loss function by Vapnik replaces the inner product operation in high-dimensional space.

This method is then applied to the regression problem, leading to the proposed support vector regression (SVR) model [19], whose structure is outlined as follows:

$$y = f(x) = w^T \varphi(x) + b \quad (4)$$

where y is the explained variable, x is the explanatory variable, $\varphi(x)$ denotes the nonlinear function that maps the explanatory variable in the original input space to some higher dimensional space, w^T is the vector of weights, and b is the bias.

According to the structural risk minimization principle, the objective function can be minimized to:

$$\begin{cases} \min \left[\frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\delta_i + \delta_i^*) \right] \\ \text{s.t.} \begin{cases} y_i - (w \cdot \varphi(x_i) + b) \leq \epsilon + \delta_i^* \\ (w \cdot \varphi(x_i) + b) - y_i \leq \epsilon + \delta_i \\ \delta_i \geq 0, \delta_i^* \geq 0 \quad i = 1, 2, \dots, l \end{cases} \end{cases} \quad (5)$$

In Equation (5), C denotes the penalty factor, ϵ is the insensitive loss function, δ_i and δ_i^* are the slack variables, and l is the sample size. After obtaining the above model, the Lagrange function can be introduced, after which the function is converted to dyadic form to obtain the final regression function as described in [20]:

$$y = f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) K(x_i, x_j) + b \quad (6)$$

Of these, α_i and α_i^* are the Lagrange multipliers, and $K(x_i, x_j)$ is the kernel function. There are three commonly used kernel functions—the polynomial kernel function, the Gaussian radial basis kernel function (RBF), and the multilayer perceptron kernel function. Since the independent and dependent variables are not usually in a simple linear relationship, and since the RBF kernel function can be more easily generalized and exhibits stronger learning ability [21], the RBF kernel function was selected for this paper; this can be expressed as:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|}{2\sigma^2}\right) \quad (7)$$

2.3. MIDAS-SVR Model

The MIDAS model is able to handle data of varying frequencies with no information loss. However, its ability to predict may be constrained by the selection of the weight function and the implicit linear assumptions within the model. In addition to this, the mixed-frequency sampling model exhibits heightened sensitivity to both noise and outlying data points. The SVR model is able to capture nonlinear features within the data by means of the kernel function (such as the RBF kernel), while also demonstrating good generalization, easy interpretation, and good robustness when confronted with outliers. The method can effectively solve the problem of a small number of samples and high dimensionality [22]. Compared to linear regression (LR), which is also a machine learning approach, this nonlinear method demonstrates better analysis results [23]. In this paper, the MIDAS-SVR model is presented, which retains the original information of the high-frequency data, thus mitigating any information loss caused by downsampling. It also enhances the model's ability to capture nonlinear features, thereby increasing prediction accuracy. To address the limitations of individual MIDAS and SVR models, this study first outlines a multivariate unrestricted MIDAS model. The predicted values generated from this model are subsequently used in the SVR model to capture nonlinear trends and mitigate the impact of outliers, thereby further improving prediction accuracy. Figure 1

illustrates the framework of the prediction model, while the formulation of the MIDAS-SVR model is presented below:

$$\hat{Y}_{t+h} = f(g(x_{1t}, \dots, x_{nt})) \quad (8)$$

where $\hat{Y}_{t+h}$ is the model output (i.e., the forecast value), x_{it} is the model input (i.e., the influencing factors on soybean prices), function f is shown in Equation (6), and function g is shown in Equation (3). In order to predict soybean prices using the MIDAS-SVR model, it is necessary to first estimate the required parameters (including β_0 , β_{1i} , C , ϵ , and $\frac{1}{2\sigma^2}$); these are not described in detail here due to space constraints, and can be found in the literature [24,25].

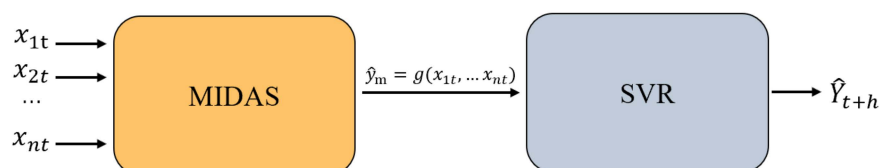


Figure 1. Framework of the MIDAS-SVR predictive model.

2.4. Evaluation Metrics

To comprehensively assess the model's predictive performance, this study employed the four key metrics of mean square error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and absolute relative error (ARE). The formulas for these metrics are defined as follows [26,27]:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (9)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (10)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (11)$$

$$ARE = \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (12)$$

where y_i is the true value and $\hat{y}_i$ is the predicted value.

3. Empirical Analysis

3.1. Analysis of Factors Influencing Soybean Prices

This paper examines the key determinants of soybean prices based on a systematic review of 33 articles published between 2012 and 2024, which were found using the search term “soybean price influence factors” on the China National Knowledge Infrastructure (CNKI). After excluding any unrelated studies, the statistical results are presented in Figure 2. This shows that most studies primarily focus on macroeconomic factors [28,29] and supply and demand dynamics [30,31], with limited investigation into weather conditions, public opinion on social media, and policy interventions.

The impact of soybean price subsidy policies on market prices operates through an indirect transmission mechanism—these policies influence soybean planting area decisions [32], which subsequently alter supply volumes and ultimately affect market prices. This process exhibits significant temporal lag effects. Furthermore, the effectiveness of such subsidies is contingent upon external factors such as agricultural technology adoption levels and farm size heterogeneity [33], which introduce additional variability in policy

outcomes. Given these confounding influences, subsidy-related variables are excluded from the feature set in this study to maintain model interpretability and predictive stability.

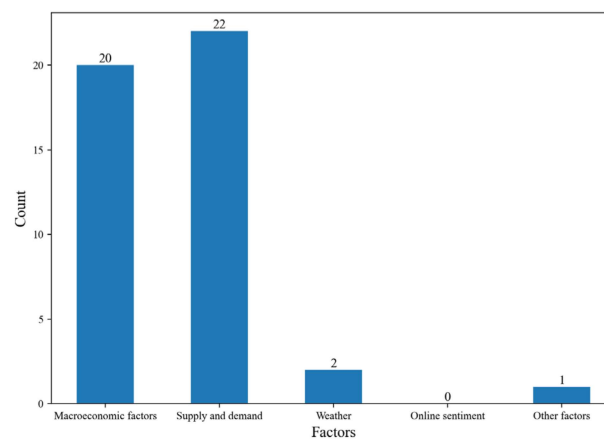


Figure 2. Literature and statistical data map of factors influencing soybean prices.

Extreme weather events create complex dynamics for crop producers by reducing local yields while simultaneously driving up prices through widespread production losses [34]. These price effects are further modulated by the El Niño Southern Oscillation (ENSO), with La Niña events exerting a particularly strong influence on both the soybean-to-corn price ratio and related hedging strategies [35].

Advances in online technology, combined with media reports and public sentiment, have influenced commodity prices [36]. Research by Li et al. developed a comprehensive list of influential factors using topic modeling, a text mining method, demonstrating that these identified factors, along with sentiment-based variables, are particularly effective in price forecasting. Their proposed framework showed superior performance in medium- and long-term forecasting over benchmark models [37]. To address these omissions, this study incorporates weather data and public opinion factors into the forecasting model. A number of potential influencing factors for soybean prices were identified, along with their corresponding measurement indices, as presented in Table 1.

Table 1. Possible influences on soybean prices.

Primary Indicator	Secondary Indicator	Source	Data Frequency	Unit
Macroeconomic factors	China Consumer Price Index (x_1)	National Bureau of Statistics	Monthly	CNY/USD
	Grain Consumer Price Index (x_2)	National Bureau of Statistics	Monthly	
	Exchange rate (x_3)	People's Bank of China	Daily	
Supply and demand	Corn price (x_4)	National Bureau of Statistics	Monthly	CNY/kg
	Soybean meal price (x_5)	Dalian Commodity Exchange	Monthly	CNY/t
	Soybean oil price (x_6)	Dalian Commodity Exchange	Monthly	CNY/t
	Diesel price (x_7)	Energy Information Administration	Monthly	USD/gal
Weather	El Niño Index (x_8)	United States National Oceanic and Atmospheric Administration	Monthly	
	Average temperature in provincial capitals of northeast China (x_9)	National Bureau of Statistics	Monthly	°C
	Average precipitation in provincial capitals of northeast China (x_{10})	National Bureau of Statistics	Monthly	°C
Online sentiment	PC-based Baidu Index for soybean price (x_{11})	Baidu Index Website	Daily	
	Mobile-based Baidu Index for soybean price (x_{12})	Baidu Index Website	Daily	
	Aggregated Baidu Index for soybean price (x_{13})	Baidu Index Website	Daily	

Note: Blank unit fields denote dimensionless quantities.

3.2. Data Sources

The target variable, soybean price, is obtained from the National Bureau of Statistics of China (NBS) at a monthly frequency. Other data sources and variables for this study are detailed in Table 1, covering the period from January 2012 to January 2024. The dataset consists of nine monthly frequency variables and four daily frequency variables. In order to ensure temporal consistency and to address data availability issues, the following standardized preprocessing procedures were applied to high-frequency data:

- Baidu Index data were normalized to 30 daily observations per month;
- USD-CNY exchange rate data (excluding weekends and holidays) were standardized to 20 observations per month;
- If the number of data points exceeded the preset threshold, any surplus data were discarded; if it fell below the threshold, any gaps were filled using data from adjacent months.

This preprocessing protocol ensures data uniformity and temporal alignment for mixed-frequency analysis. The study period was divided into two subsets:

- Training set (January 2012–January 2021): Used for model parameter estimation.
- Test set (February 2021–January 2024): Reserved for out-of-sample performance evaluation.

The temporal partitioning of our dataset follows standard practices in time series forecasting research, with training–testing splits systematically implemented within the conventional 60:40 to 80:20 ratio range [38].

3.3. Variable Selection

A multidimensional variable screening method was used in this study to identify the key variables for the soybean price forecasting model. Monthly frequency data were analyzed to identify variables with a statistically significant relationship to soybean price by calculating the Pearson and Spearman correlation coefficients, followed by a double correlation test; the results are presented in Table 2.

Table 2. Correlations Between Predictors and Soybean Prices: Pearson (Linear) and Spearman (Rank) Coefficient (February 2012–January 2021).

Indicators	Pearson Correlation Coefficient	Significance	Spearman Rank Correlation Coefficient	Significance
China Consumer Price Index (x_1)	0.367 **	0.000	0.083	0.389
Grain Consumer Price Index (x_2)	0.318 **	0.001	0.086	0.372
Corn price (x_4)	0.480 **	0.000	0.525 **	0.000
Soybean meal price (x_5)	−0.046	0.631	0.185	0.054
Soybean oil price (x_6)	−0.165	0.087	−0.054	0.574
Diesel price (x_7)	−0.178	0.064	−0.113	0.241
El Niño Index (x_8)	−0.223 **	0.020	−0.108	0.263
Average temperature in provincial capitals of northeast China (x_9)	0.114	0.237	0.015	0.875
Average precipitation in provincial capitals of northeast China (x_{10})	0.029	0.766	0.005	0.957

Note: ** denotes significance at the 0.01 level.

Table 2 demonstrates a statistically significant positive correlation between soybean prices and both the China Consumer Price Index and the Grain Consumer Price Index, consistent with conventional price transmission mechanisms whereby inflationary pressures increase production costs in agricultural inputs (e.g., fertilizers and transportation), consequently elevating commodity market prices [39]. Furthermore, the corn and soybean markets exhibit a substitution relationship on the demand side (as evidenced by feed ingredient conversion) [40]. Finally, El Niño events generally lead to warmer and drier

conditions in some regions, which can reduce soybean yields, thereby increasing prices due to supply shortages [41].

The input variables for daily frequency data were selected for optimal predictive accuracy based on the performance assessment of the univariate mixed frequency data sampling (MIDAS) model. The specific results are presented in Table 3. The final variable selection results are collectively determined based on comprehensive analysis of both Tables 2 and 3, where we retained only the optimal indicator at each frequency level. The Chinese consumer price index, corn price, El Niño index, exchange rate, and aggregated Baidu index for soybean price were identified as the key influencing factors for inclusion in the MIDAS-SVR prediction model.

Table 3. The results of the Single-Variable MIDAS Model (February 2012–January 2021).

Variables	Training Set MAPE (%)
Exchange rate (x_3)	1.57
PC-based Baidu Index for soybean price (x_{11})	1.97
Mobile-based Baidu Index for soybean price (x_{12})	2.07
Aggregated Baidu Index for soybean price (x_{13})	1.90

3.4. Prediction Results and Analysis Based on the MIDAS-SVR Model

This study seeks to predict future soybean price trends by means of a MIDAS-SVR forecasting model. Preprocessed data were used to predict prices within the R programming environment, with the detailed methodology outlined in Section 2. The model is initially trained with a training set to identify the optimal parameters for the MIDAS-SVR soybean price prediction model (refer to Table 4). Figure 3 illustrates the fitted values from the prediction model alongside the training set error metrics—the mean squared error (MSE) is 0.013, the mean absolute error (MAE) is 0.08, and the mean absolute percentage error (MAPE) is 1.36%. These results suggest that the model demonstrates high predictive accuracy.

Table 4. Parameter Settings of the MIDAS-SVR Model.

Parameter Name	Parameter Value	Parameter Name	Parameter Value
β_0	1.711	h	1
β_1	1.035	m_1	1
β_2	3.612×10^{-2}	m_2	1
β_3	-7.210×10^{-2}	m_3	1
β_4	2.147×10^{-2}	m_4	20
β_5	-5.653×10^{-5}	m_5	30
C	200	ϵ	0.1
$K(x_i, x_j)$	radial (RBF)	$\frac{1}{2\sigma^2}$	0.01

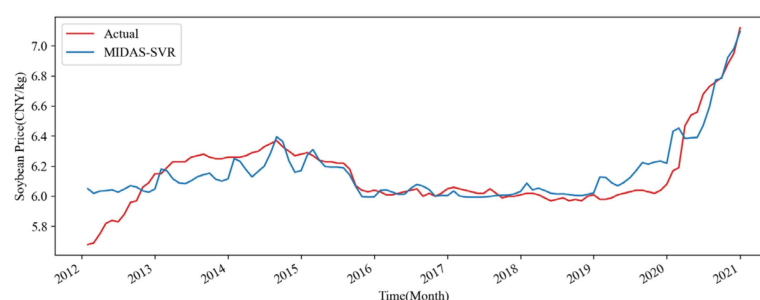


Figure 3. Training Fitting Curve for MIDAS-SVR Soybean Price Prediction Model.

The variables for corn price, consumer price index, El Niño index, exchange rate, and Baidu soybean price index are arranged in ascending order. The notation represents the coefficient for the corn price variable while indicating the multiplicative relationship between the duration of the corn price cycle and that of the soybean price cycle, among other relationships. The MIDAS-SVR model was employed to predict soybean prices for the test sample covering February 2021 to January 2024, with the results presented in Table 5 and Figure 4. The prediction accuracy presented in Table 5 exhibits an initial trend of gradual improvement, followed by a plateau. The inaccuracy encountered in 2021 is likely associated with the ongoing effects of the COVID-19 pandemic. In early 2021, the volatility of soybean market prices escalated significantly due to supply chain disruptions, demand variations, and policy changes instigated by the pandemic, resulting in the suboptimal performance of the model during initial forecasts. After that, the uncertainty impact of the pandemic on the market decreased, and the accuracy of the model improved. Beginning in November 2021, there was a marked reduction in the model's prediction error, with the relative error stabilizing at approximately 1%. This suggests that the model was better at capturing price trends during the later phases of the epidemic.

Table 5. Prediction Results and Error Analysis of the MIDAS-SVR Model on the Test Set.

Time	Actual Values	Predicted Values	ARE(%)	Time	Actual Values	Predicted Values	ARE(%)
February 2021	7.19	7.10	1.31	August 2022	7.85	7.85	0.05
March 2021	7.24	7.67	6.01	September 2022	7.89	7.88	0.17
April 2021	7.19	7.72	7.36	October 2022	7.91	7.88	0.40
May 2021	7.22	7.67	6.29	November 2022	7.87	7.89	0.28
June 2021	7.23	7.61	5.28	December 2022	7.92	7.91	0.15
July 2021	7.27	7.59	4.42	January 2023	7.94	7.94	0.00
August 2021	7.28	7.60	4.42	February 2023	7.90	7.90	0.03
September 2021	7.27	7.58	4.26	March 2023	7.87	8.04	2.18
October 2021	7.34	7.61	3.64	April 2023	7.83	7.95	1.54
November 2021	7.45	7.48	0.40	May 2023	7.81	7.88	0.91
December 2021	7.56	7.45	1.51	June 2023	7.77	7.78	0.07
January 2022	7.58	7.57	0.20	July 2023	7.78	7.68	1.26
February 2022	7.62	7.56	0.72	August 2023	7.80	7.68	1.55
March 2022	7.71	7.71	0.04	September 2023	7.82	7.72	1.32
April 2022	7.74	7.70	0.46	October 2023	7.74	7.80	0.74
May 2022	7.79	7.81	0.29	November 2023	7.72	7.75	0.34
June 2022	7.82	7.81	0.12	December 2023	7.69	7.59	1.25
July 2022	7.84	7.79	0.70	January 2024	7.63	7.49	1.84

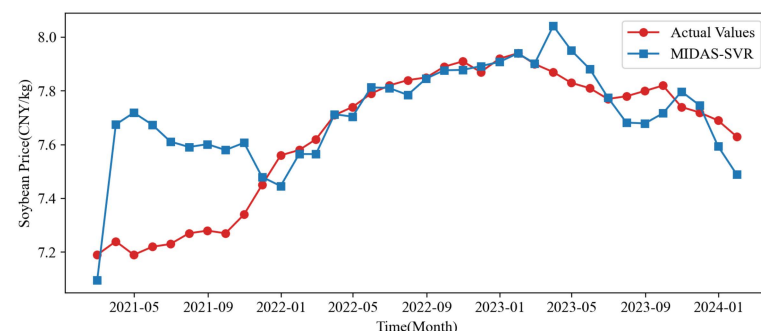


Figure 4. Prediction Results of the MIDAS-SVR Model for Soybean Prices (Test Set).

3.5. Comparative Analysis

This study aims to rigorously evaluate the performance of the MIDAS-SVR model in forecasting soybean prices, particularly its capacity to analyze the intricate relationship between high-frequency factors and soybean prices. The MIDAS-SVR model is compared with five alternative models—a mixed-frequency data sampling (MIDAS) model, a traditional support vector regression (SVR) model using single-frequency data inputs (Baidu Index for Soybean Prices and average exchange rate), and a multilayer perceptron (MIDAS-MLP)

model based on mixed-frequency data. The SVR model exclusively uses low-frequency data, specifically the monthly averages of the Baidu index and exchange rate variables. The MIDAS-MLP model parallels the MIDAS-SVR model by using predicted values from the MIDAS model as inputs for the MLP model; this is capable of directly processing mixed-frequency data, including high-frequency Baidu index and low-frequency economic indicators. We also compared the MIDAS-SVR model against standard time series models (auto-ARIMA and ETS). Figure 5 displays the fitted values for each model, showing how the six fitting curves align with trends observed in the actual sample curves.



Figure 5. Training Fitting Curves for the Four Prediction Models.

Figure 5 illustrates that the MIDAS model is less suitable for capturing soybean price trends than the other five models. In early 2016, mid-2017, and throughout 2019 and 2020, the fitted values of the MIDAS model exhibited significant deviations. The MIDAS-SVR and SVR models demonstrated a superior fit during that time period. Among the compared models, auto-ARIMA and ETS exhibit the highest fitting accuracy.

Figures 6 and 7 present the residual histograms and residual boxplots for the six models, respectively. Figure 6 illustrates that the residuals of both the MIDAS model and the MIDAS-MLP model exhibit a distinctly left-skewed distribution, suggesting reduced stability. Figure 7 illustrates that the error means for each model approach zero, suggesting reduced bias within the training set. Additional analysis of the error distributions indicates that the auto-ARIMA and ETS models exhibit more concentrated error distributions with reduced fluctuation ranges, demonstrating better predictive stability. The error distributions of the MIDAS model and the MIDAS-MLP model exhibit greater dispersion, suggesting a higher level of uncertainty in their prediction outcomes.

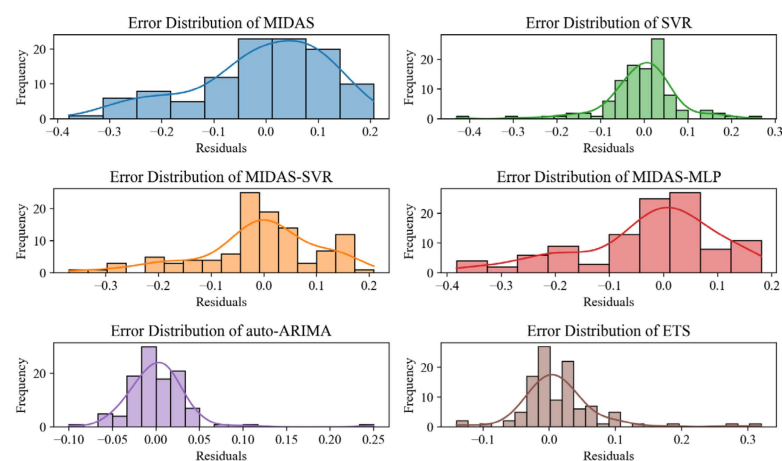


Figure 6. Residual Analysis of the Four Models on the Training Set.

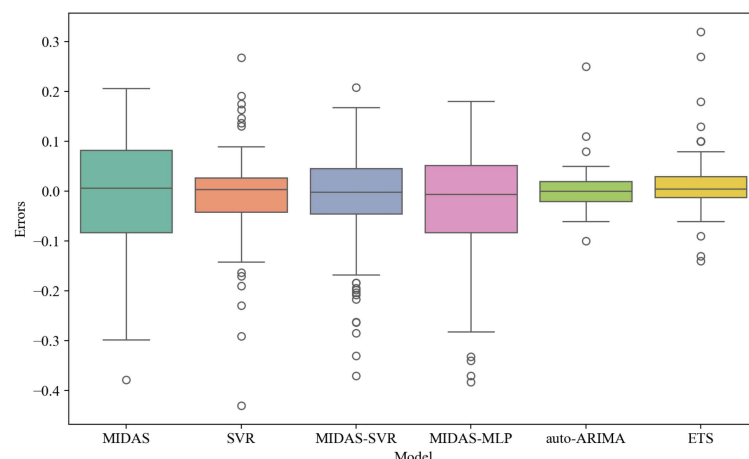


Figure 7. Boxplots of Residuals from the Six Models.

This study evaluated the forecasting performance of the MIDAS-SVR model by predicting soybean prices from February 2021 to January 2024 (test sample) using six different models. The results are presented in Table 6 and Figure 8.

Table 6. Prediction Errors of the Different Models (Test Set, February 2021–January 2024).

Model	MAPE	MAE	MSE
MIDAS	4.16%	0.32	0.13
SVR	2.36%	0.18	0.05
MIDAS-SVR	1.71%	0.13	0.04
MIDAS-MLP	2.02%	0.15	0.04
ETS	6.90%	0.53	0.50
Auto-ARIMA	14.41%	1.12	1.74

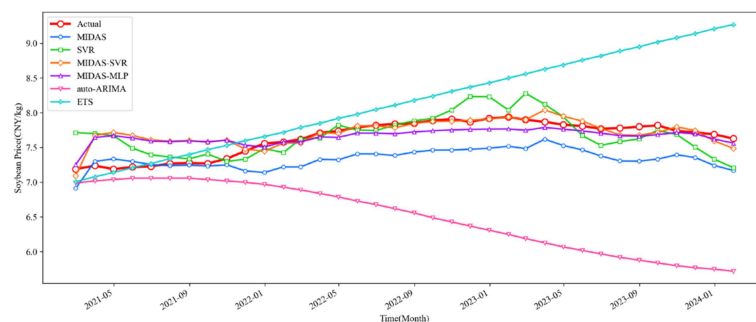


Figure 8. Prediction Results of the Six Models (Test Set).

The empirical results demonstrate significant performance limitations in traditional time series models, with ETS achieving 6.90% MAPE (MAE 0.50 and MSE 0.53) and auto-ARIMA performing substantially worse at 14.41% MAPE (MAE 1.74 and MSE 1.12) in out-of-sample testing. These limitations stem from fundamental structural constraints—while their linearity and stationarity assumptions provide protection against overfitting, they inherently restrict the models' ability to capture complex nonlinear patterns. In contrast, our MIDAS-SVR framework achieves superior predictive accuracy (1.71% MAPE, MAE 0.04, and MSE 0.13), demonstrating reductions of 75.2%, 92.0%, and 75.5%, respectively, compared to ETS on these metrics. This performance advantage, extending to 88.1%, 97.7%, and 88.4% improvements over auto-ARIMA, validates that machine learning-enhanced mixed-frequency modeling can simultaneously achieve greater accuracy while maintaining robustness against overfitting.

Furthermore, comparative analysis reveals that MIDAS-SVR outperforms three alternative hybrid approaches—conventional MIDAS (4.16% MAPE), standard SVR (2.36% MAPE), and MIDAS-MLP (2.02% MAPE)—demonstrating the unique advantages of our proposed architecture in handling mixed-frequency data while avoiding overfitting.

3.6. Robustness Tests

Predictive robustness evaluates the stability and reliability of a predictive model's outputs under varying conditions or environmental changes. This analysis seeks to confirm that the model consistently delivers accurate and reliable predictions in real-world applications, while also maintaining robust performance when confronted with data fluctuations, outliers, or unrealistic model assumptions. The paper employs two methods to assess the robustness of the model—a reduced training period (excluding the 2012 data) and an increased number of independent variables.

The MIDAS-SVR model was retrained by shortening the training set period (January 2013–January 2021), after which its predictive performance was assessed. Table 7 indicates that the set error, measured by MAPE, is 2.15%. In addition, the mean absolute error (MAE) exhibits an increase of 0.03 units, while the mean square error (MSE) shows a relative rise of 0.01 units. This suggests that a reduction in the training set data results in inadequate historical information being captured by the model, potentially diminishing its capacity for generalization. The error remains within an acceptable range.

Table 7. Robustness Test.

Plan	MAPE	MAE	MSE
Original MIDAS-SVR model	1.71%	0.13	0.04
Incorporating soybean meal price variables	1.72%	0.13	0.04
Reducing training time (January 2013–January 2021)	2.15%	0.16	0.05

This paper introduces soybean meal price as an explanatory variable to address potential omitted variable bias in the existing literature on factors influencing soybean prices. Upon incorporating the soybean meal price variable, the MIDAS-SVR model was re-evaluated, revealing that the mean relative error (MAPE) of the test set increased by a mere 0.01% (refer to Table 7). This suggests that the inclusion of explanatory variables slightly decreases the fitting accuracy on the training set while leaving prediction accuracy largely unaffected. The model presented in this paper successfully passed the robustness test, indicating that the research conclusions are very reliable.

4. Results and Discussion

Soybean prices are influenced by a variety of factors, including both macroeconomic data, which is typically reported on a quarterly basis, and low-frequency data, which is often available on a monthly basis. This paper employs a mixed-frequency data sampling (MIDAS) model to integrate daily data, such as weather and investor sentiment, with varying frequency sample data, all within a unified prediction framework. Our approach extends the work of Wang et al. (2023) by incorporating machine learning techniques to address known nonlinearity challenges in standard MIDAS applications [11]. Given the complexity and nonlinear nature of soybean price fluctuations, this study leverages the strengths of machine learning in handling intricate data. By fully exploring the nonlinear and discontinuous aspects of the data, the research combines machine learning with a mixed-frequency data sampling model to develop a mixed-frequency support vector regression (MIDAS-SVR) prediction model for soybean prices. Empirical research on soybean price data in China from January 2012 to January 2024 indicates that the MIDAS-

SVR model demonstrates optimal predictive performance and improved stability, leading to the following key conclusions:

- There are four key dimensions that influence soybean prices: macroeconomic factors, supply and demand, weather, and public sentiment. Notably, the Consumer Price Index (CPI), corn prices, the El Niño Index, exchange rates, and the Baidu Index exhibit a significant correlation with fluctuations in soybean prices (p -values < 0.05). These findings align with Ferreira et al.'s (2022) demonstration of weather impacts on soybean markets [35].
- The MIDAS-SVR model improves prediction accuracy by using the information contained in high-frequency data, thus improving on traditional methods that merely simplify high-frequency data to single-frequency data through direct averaging.
- The MIDAS-SVR model demonstrates superior prediction accuracy and stability compared to other models, making it very suitable for capturing soybean price trends. The results obtained in this paper validate the MIDAS-SVR prediction model and offer new methodological support for soybean price prediction, opening up a wide range of theoretical and practical applications.

It should be noted that the current study has two main limitations: The model validation was conducted on a single commodity time series, and its generalizability to other agricultural markets requires further investigation. China-focused data may require calibration for global markets. Future research will expand the validation to multiple commodity markets and explore optimization techniques for practical implementation.

Author Contributions: Conceptualization, X.L. and D.Z.; methodology, X.L.; software, X.L.; validation, X.L.; formal analysis, X.L.; investigation, X.L., W.Z. and Z.G.; resources, X.L., W.Z. and Z.G.; data curation, X.L., W.Z. and Z.G.; writing—original draft preparation, X.L.; writing—review and editing, X.L., D.Z. and K.M.; visualization, X.L.; supervision, D.Z. and K.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Jiangsu Province College Student Innovation and Entrepreneurship Training Program (Grant No. 202410307248Y).

Data Availability Statement: The data regarding soybean prices and their influencing factors are available at: <https://data.stats.gov.cn/easyquery.htm?cn=A01> (accessed on 16 May 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Central Committee of the Communist Party of China (CPC) and State Council. Opinions on Learning and Applying the Experience of the 'Project of Demonstrating 1000 Villages and Renovating 10,000 Villages' to Effectively Promote Comprehensive Rural Revitalization. Central Document No. 1, 1 January 2024. Available online: https://www.gov.cn/gongbao/2024/issue_11186/202402/content_6934551.html (accessed on 3 January 2025).
2. Zhu, J.; Fan, Y.D.; Xu, Y. Soybean price prediction in China based on modified GM (1,1) model. *Soybean Sci.* **2016**, *35*, 315–319.
3. Xiong, T.; Bao, Y.K. Soybean future prices forecasting based on dynamic model averaging. *Chin. J. Manag. Sci.* **2020**, *28*, 79–88.
4. Yang, J.; Zhang, D.B.; Fang, J.F.; Li, P.H. Domestic soybean price forecast based on EEMD and support vector regression. *Guangdong Agric. Sci.* **2019**, *46*, 134–140.
5. Zhang, D.Q.; Zang, G.M.; Li, J.; Ma, K.P.; Liu, H. Prediction of soybean price in China using QR-RBF neural network model. *Comput. Electron. Agric.* **2018**, *154*, 10–17. [CrossRef]
6. Fan, J.M.; Liu, H.J.; Hu, Y.R. Soybean future prices forecasting based on LSTM deep learning. *Prices Mon.* **2021**, *2*, 7–15.
7. Ghysels, E.; Santa-Clara, P.; Valkanov, R. *The MIDAS Touch: Mixed Data Sampling Regressions*; Mimeo: Chapel Hill, NC, USA, 2004.
8. Wu, Q. Research on the application of mixed frequency model in macroeconomic forecasting in China. *Price Theory Pract.* **2024**, *2*, 28–34+108.
9. Guo, Y.L.; Ma, F. Forecasting the Chinese gold futures market volatility using Markov-switching regime and mixed data sampling model. *Chin. J. Manag. Sci.* **2024**, *32*, 13–22.

10. Cai, Y.; Tang, Z.P.; Wu, J.C.; Du, X.X.; Chen, K.J. Research on the application of the GWO-SVR algorithm in the prediction of reverse mixed data in stock market and investment strategy. *Chin. J. Manag. Sci.* **2024**, *32*, 73–80.
11. Wang, L.; Wu, R.; Ma, W.C.; Xu, W.J. Examining the volatility of soybean market in the MIDAS framework: The importance of bagging-based weather information. *Int. Rev. Financ. Anal.* **2023**, *89*, 102720. [[CrossRef](#)]
12. Yang, Y.; Rong, H.L.; Cheng, G.X.; Gao, H. Soybean futures responses to meteorological disaster risk—Empirical evidence from the Chicago board of trade. *Financ. Res. Lett.* **2025**, *78*, 106904. [[CrossRef](#)]
13. Li, B.; Shao, X.Y.; Li, Y.Y. Research on machine learning-driven quantamental investing. *China Ind. Econ.* **2019**, *8*, 61–79.
14. Jiang, F.; Zhang, W.Y. Application of machine learning methods in economic research. *Stat. Decis.* **2022**, *38*, 43–49.
15. Liu, T.; Yang, W.M.; Hu, R.T. Empirical study on quarterly GDP forecasting using mixed-frequency data based on AIC criterion. *Stat. Theory Pract.* **2021**, *6*, 26–33.
16. Liu, K.B.; Zhang, T. Short-term and inflection point forecasting of CPI by using Internet searching big data: An Empirical Study Based on MIDAS Model. *Contemp. Financ. Econ.* **2018**, *11*, 3–15.
17. Ding, L.L.; Sun, W.X.; Han, M.; Kang, W.L. Influence of PMI index on GDP and its prediction effect in China. *Stat. Decis.* **2018**, *34*, 128–132.
18. Zhou, J.; Tang, C.Q. Research on China's macroeconomic forecasting mechanism mixed frequency weighted sampling model. *Econ. Probl.* **2018**, *6*, 1–5+85.
19. Vapnik, V.; Levin, E.; Lecun, Y. Measuring the VC-dimension of a learning machine. *Neural Comput.* **1994**, *6*, 851–876. [[CrossRef](#)]
20. Sun, Q.Y.; Liu, J.Q.; Liu, Y.; Wu, Q.X. Stock Price prediction model based on SVR with parameters optimized by improved GA. *Comput. Syst. Appl.* **2015**, *24*, 29–34.
21. Zhou, C.; Bai, B.; Ye, N. Reliability prediction of engineering system based on adaptive particle swarm optimization support vector regression. *J. Mech. Eng.* **2023**, *59*, 328–338.
22. Zhang, R.; Liu, Y. Research on development and application of support vector machine—Transformer fault diagnosis. In Proceedings of the ISBDAI '18: International Symposium on Big Data and Artificial Intelligence, Hong Kong, 29–30 December 2018; pp. 262–268.
23. Aung, Z.; Mihailov, I.S.; Aung, Y.T. Models and data mining algorithms for solving classification problems. In Proceedings of the 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA 2019), Lipetsk, Russia, 20–22 November 2019; pp. 532–536.
24. Wang, W.G.; Yu, Y. Short-term prediction of quarterly GDP in China based on MIDAS regression models. *J. Quant. Tech. Econ.* **2023**, *59*, 328–338.
25. Liu, J.X. Support vector regression based on grid search hyperparameter optimization. *Sci. Technol. Innov.* **2022**, *13*, 71–74.
26. Xia, M.S.; Jiang, L.L. China's consumer confidence index forecast based on deep network CNN-LSTM model. *Stat. Decis.* **2021**, *37*, 21–26.
27. Zhang, X.; Du, J.L. PM2.5 concentration prediction based on improved PSO-GA-BP. *Comput. Eng. Des.* **2019**, *40*, 1718–1723.
28. Zha, T.J. The possibility of China gaining pricing power in soybean futures: A principal component analysis of factors influencing domestic soybean prices. *Financ. Theory Pract.* **2016**, *1*, 37–41.
29. Liu, H.; Zhang, D.Q. Analysis on influencing factors of domestic soybean price based on quantile regression. *Soybean Sci.* **2014**, *33*, 759–763.
30. Fan, Z.; Ma, K.P.; Jiang, S.J.; Shi, N. Influence factors analysis and price prediction of soybean in China based on improved GM (1,N) model. *Soybean Sci.* **2016**, *35*, 847–852.
31. Gao, L.; Zhang, D.Q.; Ye, F.R.; Huang, N. Analysis on influencing factors of domestic soybean price based on symbolic regression. *Soybean Sci.* **2017**, *36*, 782–788.
32. Guo, B.S.; Xin, L.Q. Application of Numerical Methods in the Study of the Impact of Agricultural Subsidy Policies on the Development of China's Soybean Industry. *J. Comb. Math. Comb. Comput.* **2025**, *127a*, 9219–9237.
33. Guo, S.; Lv, X.; Hu, X. Farmers' land allocation responses to the soybean rejuvenation plan: Evidence from “typical farm” in Jilin, China. *China Agric. Econ. Rev.* **2021**, *13*, 705–719. [[CrossRef](#)]
34. Lee, S. Effects of extreme heat events on crop revenues for U.S. corn and soybeans. *Am. J. Agric. Econ.* **2025**, *1*–28. [[CrossRef](#)]
35. Ferreira, G.L.M.; Tonin, J.M.; Alves, A.F. Impacts of El Niño southern oscillation on hedge strategies for Brazilian corn and soybean futures contracts. *Rev. De Econ. E Sociol. Rural.* **2022**, *60*, e250643. [[CrossRef](#)]
36. Xu, L.A.; Zhao, C.S.; Song, Z.Y. Crude oil price forecasting with online news topic distribution and news sentiment classified by topics. *China J. Econom.* **2023**, *3*, 443–463.
37. Li, J.; Li, G.; Liu, M.; Zhu, X.; Wei, L. A novel text-based framework for forecasting agricultural futures using massive online news headlines. *Int. J. Forecast.* **2022**, *38*, 35–50. [[CrossRef](#)]
38. Bichri, H.; Chergui, A.; Hain, M. Investigating the impact of train/test split ratio on the performance of pre-trained models with custom datasets. *Int. J. Adv. Comput. Sci. Appl.* **2024**, *15*, 331–339. [[CrossRef](#)]

39. Rozi, F.; Subagio, H.; Elisabeth, D.A.A.; Mufidah, L.; Saeri, M.; Supriyadi; Burhansyah, R.; Kilmanun, J.C.; Krisdiana, R.; Hanif, Z.; et al. Indonesian foodstuffs in facing global food crisis: Economic aspects of soybean farming. *J. Agric. Food Res.* **2025**, *19*, 101669.
40. Caldarelli, C.E.; Bacchi, M.R.P. Factors influencing the price of corn in Brazil. *Nova Econ.* **2012**, *22*, 141–164. [[CrossRef](#)]
41. Sun, T.T.; Wu, T.; Chang, H.L.; Tanasescu, C. Global agricultural commodity market responses to extreme weather. *Econ. Res.-Ekonom. Istraživanja* **2023**, *36*, 2186913. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.