Hindawi
Mathematical Problems in Engineering
Volume 2021, Article ID 3744320, 15 pages
https://doi.org/10.1155/2021/3744320

Research Article

Prediction of House Price Index Based on Bagging Integrated WOA-SVR Model

Xiang Wang,[1] **Shen Gao**,[2] **Shiyu Zhou**,[3] **Yibin Guo**,[1] **Yonghui Duan**,[2] **and Daqing Wu**[4]

[1]*Department of Civil Engineering, Zhengzhou University of Aeronautics, No. 15, Wenyuan West Road, Zhengdong New District, Zhengzhou 450015, China*
[2]*Department of Civil Engineering, Henan University of Technology, No. 100, Lianhua Street, Gaoxin District, Zhengzhou 450001, China*
[3]*Department of Management, Shanghai University, No. 99 Shangda Road, Baoshan District, Shanghai 200444, China*
[4]*Department of Economics and Management, Shanghai Ocean University, No. 999, Huchenghuan Road, Pudong New District, Shanghai 201306, China*

Correspondence should be addressed to Xiang Wang; shanghaiwx1976@126.com

Received 26 August 2021; Revised 2 October 2021; Accepted 6 October 2021; Published 29 October 2021

Academic Editor: Juan Frausto-Solis

Aiming at the shortcomings of a single machine learning model with low model prediction accuracy and insufficient generalization ability in house price index prediction, a whale algorithm optimized support vector regression model based on bagging ensemble learning method is proposed. Firstly, gray correlation analysis is used to obtain the main influencing factors of house prices, and the segmentation forecasting method is used to divide the data set and forecast the house prices in the coming year using the data of the past ten years. Secondly, the whale optimization algorithm is used to find the optimal parameters of the penalty factor and kernel function in the SVR model, and then, the WOA-SVR model is established. Finally, in order to further improve the model generalization capability, a bagging integration strategy is used to further integrate and optimize the WOA-SVR model. The experiments are conducted to forecast the house price indices of four regions, Beijing, Shanghai, Tianjin, and Chongqing, respectively, and the results show that the prediction accuracy of the proposed integrated model is better than the comparison model in all cases.

1. Introduction

As a pillar industry of the national economy, real estate plays an indispensable role in promoting China's economic development. In the past decade, real estate has developed at a high speed in the Chinese economic system. According to the data released by the National Bureau of Statistics, the residential investment in real estate development in China in 2020 reached RMB10.444 trillion, which rose 206.99% compared to RMB 3.402 trillion ten years ago. In recent years, in order to maintain the stability and healthy development of real estate market prices, local governments have formulated a series of policies to restrict purchasing and loaning. Therefore, strengthening long-term monitoring and accurately predicting real estate market prices is an important guideline for homebuyers to grasp the timing of home purchasing and for government departments to formulate future housing prices control policies.

Previous research on house price forecasting models has been divided into two kinds of econometric models and machine learning models. In econometric models, most models use historical prices to predict future house prices. The common models are Autoregressive Integrated Moving Average Model (ARIMA), Generalized AutoregRessive Conditional Heteroscedasticity model (GARCH), and Vector Autoregressive Models (VAR). For example, Zhao et al. [1] used the ARIMA model and multiple linear regression (MLR) model to forecast New Zealand house prices, respectively, and the experimental results showed that the ARIMA model generally outperformed the MLR model.

Hou and Qiao [2] used wavelet analysis to decompose and reconstruct the house price data of Taiyuan city and used ARIMA to forecast each component obtained from the decomposition. Wang [3] used a GARCH model to analyze the long-term equilibrium relationship between house prices and CPI in Beijing and to forecast the trend of house price changes in Beijing. Li et al. [4] used a high-dimensional sparse VAR model to study the problem of house price forecasting in 35 large and medium-sized cities in China, and the experimental results showed that the proposed model has a more streamlined structure and better forecasting results compared to the traditional VAR model. However, the above econometric models mostly use house price data as a time series for linear forecasting, which is difficult to capture the nonlinear part of the data and has poor forecasting accuracy [5].

Compared to econometric models, machine learning models are effective in mining and retaining valuable information in the data and in dealing with the nonlinear part of the data [6]. Among the most widely used machine learning models are BP neural networks and support vector regression (SVR) models. However, the BP neural network model has the defects of slow convergence, easy to fall into local extremes, and large demand for training samples and is prone to overfitting. SVR is a regression prediction model based on the principle of structural risk minimization, which performs well under small sample and high-dimensional data sample conditions with better generalization and nonlinear fitting ability compared to BP neural network models [7]. In previous studies, Ye et al. [8] proposed a wavelet Lp-norm support vector (Lp-WSVR) to improve the ability of feature selection and prediction accuracy in house price prediction problems. Han and Clemmensen [9] improved the SVR model by adding weights to the relaxation variables and experimentally demonstrated that the improved method has high accuracy and can effectively reduce the effect of outliers. Dong et al. [10] used Baidu search data as input data to predict second-hand and new house prices in 16 large and medium-sized cities in China, and the experimental results showed that the SVR model performed the best. However, when using the SVR model for regression prediction, the model prediction accuracy is usually influenced by the parameter settings, and the metaheuristic algorithm can quickly lock the optimal parameters of the model and improve the prediction accuracy of the SVR model. In a previous study, Pai and Hong [11] first used a genetic algorithm to optimize the parameters of SVM and experimentally proved that the proposed model can effectively improve prediction accuracy. Liu and Hong [12] used a particle swarm algorithm to optimize the parameters of the least squares support vector regression model and used the optimized model to forecast house price data. Tang et al. [13] used the bat algorithm to optimize the SVR model parameters and combined it with web search data to build a combined model to forecast the Beijing second-hand house price index.

In the above studies, most of them are improved from the SVR model itself or from the data selection point of view, and the essence of their predictions is all predictions of a single model. However, a single machine learning model has poor generalization ability when making predictions, which makes it difficult to achieve high-precision predictions for house price data. The ensemble learning models combine multiple homogeneous or heterogeneous base learners into one strong learner through different learning strategies, which can achieve secondary learning of the prediction results and improve the generalization ability of the prediction model. Among them, the bagging ensemble learning strategy combines the results of multiple homogeneous base learners by parallel learning, which can effectively reduce the variance of model prediction and improve the prediction accuracy and has been used in several industries. For example, Choi and Jin [14] used a bagging integration strategy for predicting solar energy output and performed bagging integration by using random forest, XGBoost, and LightGBM models as base learners, respectively. The experimental results show that all the models after bagging integration can effectively improve the prediction accuracy. Yang et al. [15] proposed a bagging integrated extreme learning machine (ELM) to predict the path of tropical cyclones in the South China Sea and demonstrated experimentally that the integrated model has high generalization ability. Jung et al. [16] used bagging integrated multilayer perceptron (MLP) for electric load forecasting.

In the light of the above analysis, this paper examines the house price index data of Beijing, Shanghai, Tianjin, and Chongqing from the following three perspectives. Firstly, we used literature analysis and gray correlation analysis to establish a house price index forecasting index system from three perspectives of macroeconomic environment, real estate industry, and market demand and used the segmented forecasting method to forecast the house price index in the coming year with the data information of the past ten years. Secondly, the SVR model is used for house price index prediction, and the whale optimization algorithm (WOA) is introduced into the parameter finding of the SVR model to establish the WOA-SVR model. Finally, the bagging ensemble learning strategy is used to integrate the WOA-SVR model for prediction in order to further improve the generalization ability and prediction accuracy of the model.

The remainder of this paper is organized as follows. Section 2 introduces the SVR, WOA, and bagging methods and then describes the prediction principles of the proposed combined model. Section 3 first introduces the indicator system of the four sets of house price index data and the idea of segmented forecasting and then introduces the parameter settings chosen for the model in this paper. Section 4 performs the model prediction performance analysis. Section 5 summarizes some conclusions.

2. Materials and Methods

2.1. Support Vector Regression Model. Support Vector Machine (SVM) [17] is a machine learning model constructed based on the principle of structural risk minimization. Support Vector Regression (SVR) model is a prediction algorithm when SVM performs regression modeling. The SVR model generalizes well and has an excellent ability to

solve nonlinear and small sample prediction problems. The basic principle of the model is shown in the following.

Given a set of training samples $N = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ with the independent variable x and the dependent variable y, the expression of the characteristic function of the SVR model can be defined as follows:

$$\begin{aligned} y &= f(x) = \omega\theta(x) + b, \\ \theta&: R^n \longrightarrow F, \omega \in F, b \in R, \end{aligned} \tag{1}$$

where $\theta(x)$ is the mapping function, ω is the weight coefficient, and b is the deviation term.

To ensure that the SVR model achieves linear differentiation on a high-dimensional space and minimizes the empirical risk, insensitive loss functions (ε) and relaxation variables ($\xi_i, \xi_i^* \geq 0$) need to be added to the model. Therefore, the objective function and constraints of the SVR model after adding the slack variables and loss functions can be defined as

$$\min \frac{1}{2}\|\omega^2\| + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{2}$$

$$\text{s.t.} \begin{cases} \left|y_i - \omega\theta(x) - b\right| \leq \varepsilon + \xi_i, \\ \xi_i, \xi_i^* \geq 0, \end{cases} \tag{3}$$

where C is the penalty factor, which is usually used to regulate both the complexity of the model and the empirical risk,ξ_i and ξ_i^* are the slack variables, and ε is the insensitivity factor.

After introducing the Lagrange multipliers α_i and α_i^*, the SVR function model at this point can be represented as follows:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)k\left(x_i, x_j\right) + b, \tag{4}$$

where $k(x_i, x_j)$ is the kernel function. Among the many kernel functions of the SVR model, the radial basis kernel function (RBF) is simple to calculate and has a better anti-interference ability for the noise present in the data, so the RBF kernel function is used in this paper, and the mathematical expression can be expressed as follows:

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}\right), \quad \sigma > 0, \tag{5}$$

where σ is the parameter of the kernel function. In practical engineering applications, the prediction performance of the SVR model is often closely related to the penalty factor C and the RBF kernel function parameter σ. Usually, C is used to adjust the model complexity, and it is easy to generate high model prediction error when C is not taken as appropriate. σ relates to the complexity of the sample mapping feature space and also affects the model prediction accuracy. Therefore, a reasonable choice of penalty factor and RBF kernel function parameters is essential for building the SVR model.

2.2. The Whale Optimization Algorithm. The whale optimization algorithm (WOA), first proposed in 2016 [18], is a population intelligence algorithm that evolves to simulate the natural hunting behavior of humpback whales with "spiral bubble nets." The algorithm has the advantages of a few adjustable parameters, simple operation, and strong search capability and has been used in several industries. In the WOA, the position of each whale represents a candidate solution of the objective function, and the candidate solution is updated to find the optimal solution through three stages: encircling prey, spiral search, and random search, until the global optimal solution is found. The following is a detailed description of the three location update phases.

2.2.1. Encircling Prey. In the prey encirclement phase, the whales first share the information about the location of the prey searched and then approach the whale closest to the current prey, gradually shrinking the encirclement. The corresponding whale position update equation in this phase is

$$\begin{cases} X(t+1) = X^*(t) - A \cdot D, \\ D = \left|C \cdot X^*(t) - X(t)\right|, \end{cases} \tag{6}$$

where t is the current number of iterations of the algorithm; X is the whale position; $X^*(t)$ is the current global optimal whale position; A and C are the coefficient matrices, which are calculated as

$$\begin{cases} A = 2a \cdot r_1 - a, \\ C = 2r_2, \\ a = \dfrac{2 - 2t}{t_{\max}}, \end{cases} \tag{7}$$

where a is the convergence factor that decreases linearly from 2 to 0, r_1 and r_2 are random numbers between the ranges [0,1], and $t_{\max}$ is the maximum number of iterations of the algorithm.

2.2.2. Spiral Search. When searching for prey, the whale uses a spiral upward to slowly approach the prey, and the expression for this phase is

$$\begin{cases} X(t+1) = X^*(t) + D \cdot e^{bl}\cos(2\pi l), \\ D = \left|C \cdot X^*(t) - X(t)\right|, \end{cases} \tag{8}$$

where b is a constant controlling the shape of the spiral, and l is a random number uniformly distributed between [−1,1].

During the humpback whale's rotational search, encircling the prey is performed simultaneously. Therefore, to simulate this simultaneous search, the WOA assumes a probability of 0.5 for each of these two position updating methods, with the following mathematical expression:

$$X(t+1) = \begin{cases} X^*(t) - A \cdot D, & p < 0.5, \\ X^*(t) + D \cdot e^{bl} \cos(2\pi l), & p \geq 0.5, \end{cases} \tag{9}$$

where p is a random number between [0,1].

2.2.3. Random Search. To enhance the global search capability of the algorithm, the WOA has a random search process to further expand the whale search range. When $|A| \geq 1$, the whale is outside the envelope and the whale searches randomly away from the current location of the prey. When $|A| < 1$, the whale uses a spiral envelope to update the position. The mathematical principle of random search is expressed as

$$\begin{cases} X(t+1) = X_{\text{rand}}(t) - A \cdot D^*, \\ D^* = \left| C \cdot X_{\text{rand}}(t) - X(t) \right|, \end{cases} \tag{10}$$

where $X_{\text{rand}}(t)$ is the random position of the whale.

2.3. WOA-SVR. The parameter optimization process of the WOA for the SVR model is essentially to use the penalty factor C and the kernel function parameter σ in the SVR model as the position vector in the whale algorithm and to find the optimal SVR parameters through iterative updates of the WOA until the global optimal position is found as the final parameter of the SVR model. The computational process is described as follows:

Step 1: initialize the algorithm parameters of WOA. Set the boundary of optimization-seeking parameters (ub and lb), the maximum number of iterations ($t_{\max}$), and the population size (N) in WOA;

Step 2: the position of each whale in the whale population is initialized, the mean square error function in the SVR model is used as the fitness function to calculate the fitness value results for each individual, and the position of the whale corresponding to the minimum adaptation value is selected as the optimal position in the current population. The fitness function can be defined by the following equation:

$$f(x_i) = \frac{1}{n} \sum_{j=1}^{n} \left(\theta_{i,j} - \widehat{\theta}_{i,j} \right)^2, \tag{11}$$

where x_i is the ith individual position, θ_{ij} is the jth true value in the training set, and $\widehat{\theta}_{ij}$ is the predicted value of the SVR model based on the parameters set by x_i individuals.

Step 3: update the whale position using equations (6), (8), and (10) and determine whether the updated whale position is within the position boundary or not, and if it is outside the boundary, generate the position randomly within the boundary.

Step 4: when the position update is completed, the updated whale position is brought into the fitness function to calculate the fitness result and compared with the minimum fitness result generated by the last iteration, and the whale position corresponding to the minimum fitness result among them is selected as the current global optimal position.

Step 5: repeat Step 3 and Step 4 and stop iteration when the maximum number of iterations is reached, at which time the global optimal position obtained is the optimal value of parameter C and σ.

Step 6: the optimal parameters output by the WOA are brought into the SVR model for modeling.

The detailed flow of the above steps is shown in Figure 1.

2.4. Bagging Ensemble Learning Strategies. Ensemble learning methods usually use some kind of combination to combine the prediction results of multiple homogeneous or heterogeneous models and effectively improve the model generalization ability and prediction accuracy [19]. Among them, the bagging ensemble learning strategy aims to integrate the prediction results of multiple homogeneous base learners, and the combination of this strategy can effectively improve the model generalization ability and avoid the occurrence of the overfitting phenomenon. The basic concept of the Bagging ensemble learning strategy is to randomly sample the dataset using the Bootstrap method and export a number of subtraining sets of N equal size. Then, base learners are trained separately on the subtraining sets, and finally, the prediction results of base learners are arithmetically averaged to obtain the final prediction results. The prediction process is shown in Figure 2.

2.5. Bagging Integrated WOA-SVR Model. The basic idea of integrating the WOA-SVR model is to use the WOA-SVR model as the base learner of the bagging strategy for ensemble learning, and its specific steps can be expressed as follows:

Step 1: data preprocessing. Preprocess the data and divide the data into train set and test set.

Step 2: bootstrap sampling. Divide the training set samples obtained in Step 1 into N subtraining sets of equal sample size using Bootstrap.

Step 3: fitting the subtraining set. N subtraining sets are fitted separately using the WOA-SVR model to obtain N base learners.

Step 4: the N base learners are predicted separately for the test set data, and the prediction results of each base learner are arithmetically averaged to obtain the final prediction result of the integrated WOA-SVR model.

The specific process is shown in Figure 3.

3. Data Description

3.1. Selection of Data and Index System. This experiment uses data related to four cities within China, namely, Beijing, Shanghai, Tianjin, and Chongqing. In the four data sets, the new residential sales price index is used as the forecast label data to measure house price changes. At the same time, because there are many factors influencing housing prices,

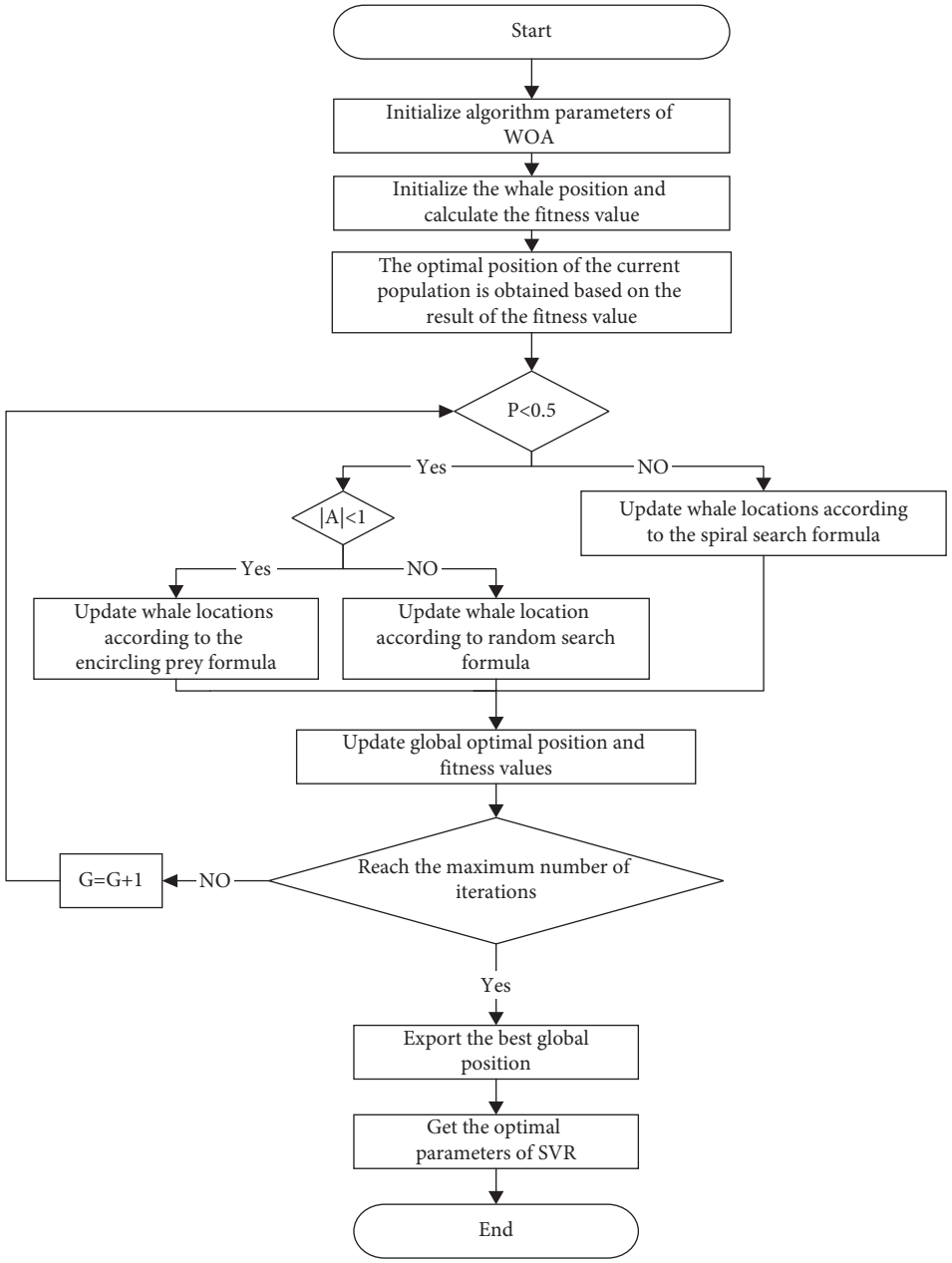


Figure 1: The flowchart of WOA-SVR.

this paper establishes the primary index system of influencing factors of house price index through literature analysis method and uses gray correlation analysis method to select the index system and find out the important indexes influencing housing prices changes.

3.1.1. Establishing a Primary Indicator System. The house price index is an indicator for quantitative analysis of the trend of real estate prices in the real estate market and is a set of indicators and analysis methods in the form of price indices to reflect the trajectory of changes in the real estate market and the current market conditions in a certain region. In this paper, we use the data of the sales price index of new residential units in 70 large and medium-sized cities published by the National Bureau of Statistics of China as the prediction labels for this experiment. Its calculation formula is defined by the following:

$$H_{i,j} = \frac{p_t^{i,j}}{p_{t-1}^{i,j}} * 100\%, \tag{12}$$

where $H_{i,j}$ is the house price index; $p_{t-1}^{i,j}$ is the average price of the jth basic classification of the ith item for period $t-1$ (last month); and $p_t^{i,j}$ is the average price in period t (this month).

As for the influencing factors of the house price index, we use the literature search method to establish the primary index system of housing prices influencing factors. Firstly, we searched CNKI journals with the keyword of "analysis of influence factors of house prices" and got 31 core journals.

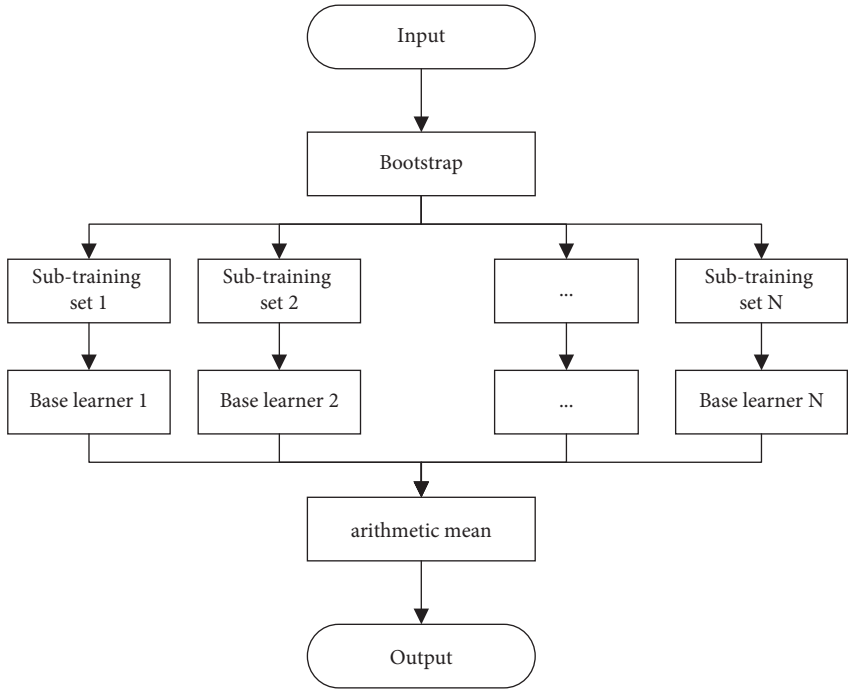


Figure 2: The flowchart of bagging ensemble learning model.

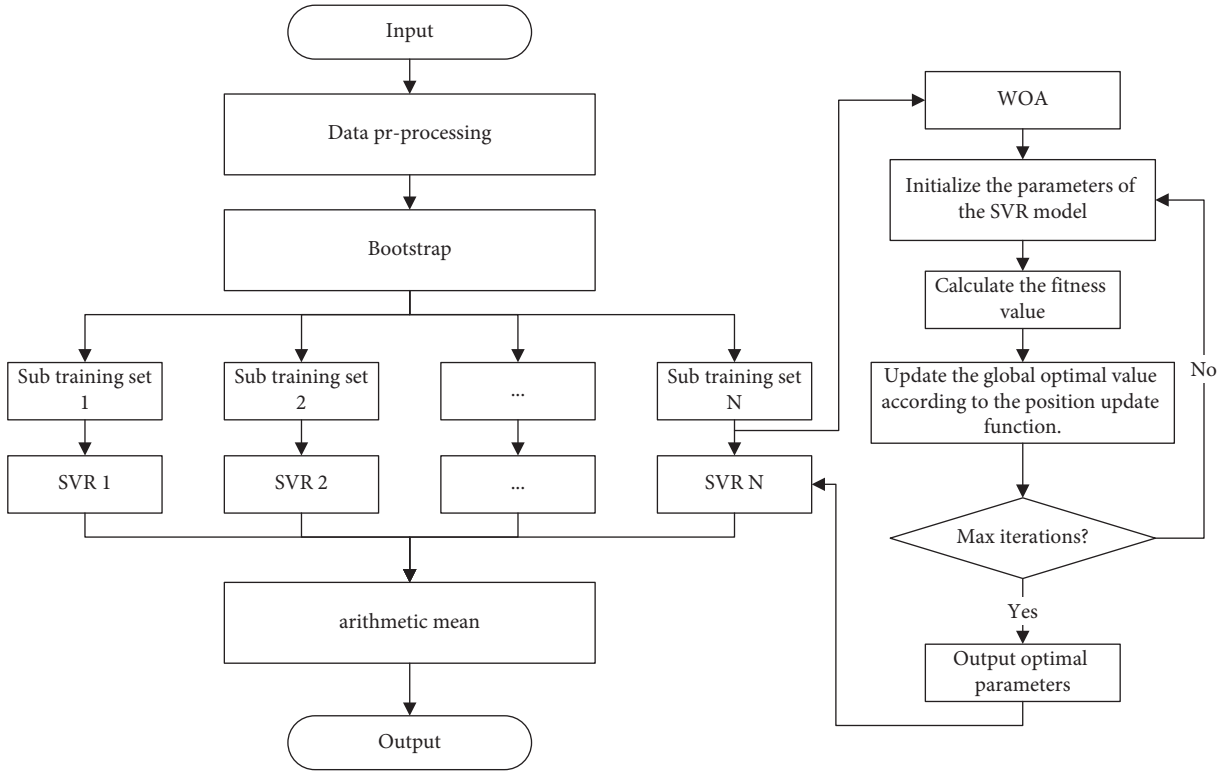


Figure 3: The flowchart of integrated WOA-SVR.

Secondly, 31 core journals were further analyzed to remove journals that did not match the theme of this study, and a total of 16 valid papers were obtained. Finally, a detailed analysis of 16 literature studies was conducted to record the factors appearing more than 5 times, and finally, the indicators of housing prices influencing factors were grouped into the following three categories: macroeconomic environment factors, real estate industry factors, and market demand factors. The primary influencing factors are shown in Table 1.

3.1.2. Grey Relation Analysis. Grey Relation Analysis (GRA) [20] is a method that measures the degree of association of the comparison series to the reference series by means of gray correlation. When the comparative series is more correlated, the series has a strong correlation with the reference series, and vice versa, the correlation is weak. The GRA process is as follows:

Step 1: use the house price index data $Y_i = \{y_{i1}, y_{i2}, \ldots, y_{in}\}$ $(i = 1, 2, \ldots, n)$ as the reference series, and the panel data of house price influences $X_j = \{x_{j1}, x_{j2}, \ldots, x_{jn}\}$ $(j = 1, 2, \ldots, n)$ as the comparison series.

Step 2: normalization of the above sequences using the min-max method.

Step 3: the gray correlation coefficients are calculated for each reference sequence and the comparison sequence. The formula is expressed as follows:

$$\xi_{ij}(k) = \frac{\min_i \min_k \left|x_j(k) - y_i(k)\right| + \rho \cdot \max_i \max_k \left|x_j(k) - y_i(k)\right|}{\left|x_j(k) - y_i(k)\right| + \rho \cdot \max_i \max_k \left|x_j(k) - y_i(k)\right|}, \tag{13}$$

where $k = 1, 2, \ldots, n$; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, n$; ρ is the resolution coefficient and takes the value range of (0,1); in this paper, we take $\rho = 0.5$.

Step 4: calculate the correlation. The formula is

$$r_{ij} = \frac{1}{n} \sum_{k=1}^{n} \xi_{ij}(k), \tag{14}$$

where r_{ij} denotes the gray correlation between the comparison sequence and the reference sequence, and when the correlation result is closer to 1, it indicates that the comparison sequence has more influence on the reference sequence.

The GRA result of the four cities' house price index data and the panel data of influencing factors is shown in Table 2. By analyzing the GRA results of the four cities in the table and removing the indicators whose correlation is less than 0.7, the final impact factor indicators obtained for each city are shown in Table 3.

3.2. Data Sources. In this paper, we selected monthly data from January 2006 to December 2020 in Beijing, Shanghai, Tianjin, and Chongqing for the experiment. The indicator data in Table 3 are published on the website of the National Bureau of Statistics, except for the ERateIndex, which is obtained from the wind database. Since the population is annual data, and GDP and income are quarterly data, Eviews10 software was used to convert the above three data to monthly. GDP, Investmentcom, and Areacom are affected by seasonal factors, and we choose the X-12 method for seasonal adjustment.

Both the HPI and CPI are month-over-month scaled indices that only reflect the ratio of price changes in the current month to those in the previous month and do not reflect the long-term overall economic trend. By consulting the statistical rules of the National Bureau of Statistics for the two kinds of data, we choose equation (12) to transform the two kinds of data. Taking the data of HPI in Beijing as an example, assuming that the statistical price in January 2006 is 100 ¥/m^2 and the HPI announced in February 2006 is 101.3, bringing it into equation (12) for the calculation to get the price of 101.3 ¥/m^2 in February 2006. The value of HPI in March 2006 is 100.6, so the transformed March price is 101.9 ¥/m^2. And so, the price index is transformed, and when the forecast is completed, it is restored back to the HPI according to equation (12). The data before and after the transformation are shown in Figure 4.

3.3. Prediction Method. In this paper, a segmented prediction method [21] is used for prediction. The overall idea is to train the model with a sample of data from the past ten years and predict the HPI in the coming year. Specifically, the first stage trains the model with data from January 2006 to December 2015 and predicts the HPI from January 2016 to December 2016. And so, the data set is finally divided into 5 segments and the last 5 years of data are predicted. The specific prediction process and the detailed time interval division of the data set are shown in Figure 5 and Table 4.

After dividing the above five prediction stages, the data set for each stage is predicted using a rolling prediction method. The basic idea is that assuming that the value of x_t at moment t is related to the data of the previous d moments, the input vector for x_t is $\{x_{t-1}, x_{t-2}, \ldots, x_{t-d}\}$, and the final prediction can be obtained by the mapping relation $f: R_d \longrightarrow R$. Its procedure is described as follows:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \\ x_2 & x_3 & \cdots & x_{d+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_{m+1} & \cdots & x_{d+m-1} \end{bmatrix}, \quad Y = \begin{bmatrix} x_{d+1} \\ x_{d+2} \\ \cdots \\ x_{d+m} \end{bmatrix}, \tag{15}$$

where X is the input vector, Y is the output vector, and d is the spatial dimension of the input vector. In this paper, we take $d = 3$, which means that we use the data of the past quarter to predict the house price of the future month.

Table 1: Primary indicators of factors influencing the house price index.

Category	Name of features	Description
Macroeconomic environment factors	LInterestRate	Interest rates for medium- and long-term loans (5 years or above)
	GDP	Gross domestic product
	CPI	Consumer price index
	ERateIndex	Exchange rate index
	M2	Money supply
Real estate industry factors	Areacom	Completed area of commercial houses
	Investmentcom	Investment in commercial housing development
	HPI_pr	Prior period house price index
Market demand factors	Income	Per capita disposable annual income of urban households
	Population	Total population at the end of the year

Table 2: GRA results.

Factors	Beijing	Shanghai	Tianjin	Chongqing
LInterestRate	0.54	0.52	0.55	0.59
GDP	0.88	0.79	0.79	0.87
CPI	0.88	0.83	0.88	0.85
ERateIndex	0.81	0.76	0.81	0.79
M2	0.92	0.87	0.90	0.87
Areacom	0.6	0.66	0.81	0.76
Investmentcom	0.81	0.85	0.89	0.81
HPI_pr	0.99	0.98	0.99	0.99
Income	0.84	0.82	0.86	0.88
Population	0.88	0.79	0.86	0.84

Table 3: Final index system of the house price index.

Category	Beijing	Shanghai	Tianjin	Chongqing
Macroeconomic environment factors	GDP	GDP	GDP	GDP
	CPI	CPI	CPI	CPI
	ERateIndex	ERateIndex	ERateIndex	ERateIndex
	M2	M2	M2	M2
Real estate industry factors	Investmentcom	Investmentcom	Investmentcom	Investmentcom
	HPI_pr	HPI_pr	HPI_pr	HPI_pr
			Areacom	Areacom
Market demand factors	Income	Income	Income	Income
	Population	Population	Population	Population

3.4. Experiment Preparation

3.4.1. Data Preprocessing. In this paper, we use the panel data of housing price influencing factors in Table 3 as input features and the transformed HPI data as output labels. And to eliminate the prediction errors due to the different data magnitudes, all predicted data are normalized by using

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{16}$$

where x^* is the normalized data value, x is the input data, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the input data.

3.4.2. Performance Measurement. To test the predictive ability of the model, mean square error (MSE) and mean absolute error (MAE) were selected to evaluate the performance of prediction results in this experiment, as follows:

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2, \\ \mathrm{MAE} &= \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y}_i\right|, \end{aligned} \tag{17}$$

where n represents the length of the data set, and y_i and $\widehat{y}_i$ represent the true and predicted values in the test set, respectively.

3.4.3. Diebold–Mariano Test. MSE and MAE can be used as evaluation criteria for model prediction performance, but the prediction error between the two compared models may not be significant when performing regression prediction. Therefore, the Diebold–Mariano (DM) test is used in this paper to verify the statistical error in the out-of-sample prediction accuracy of the two comparison models. The

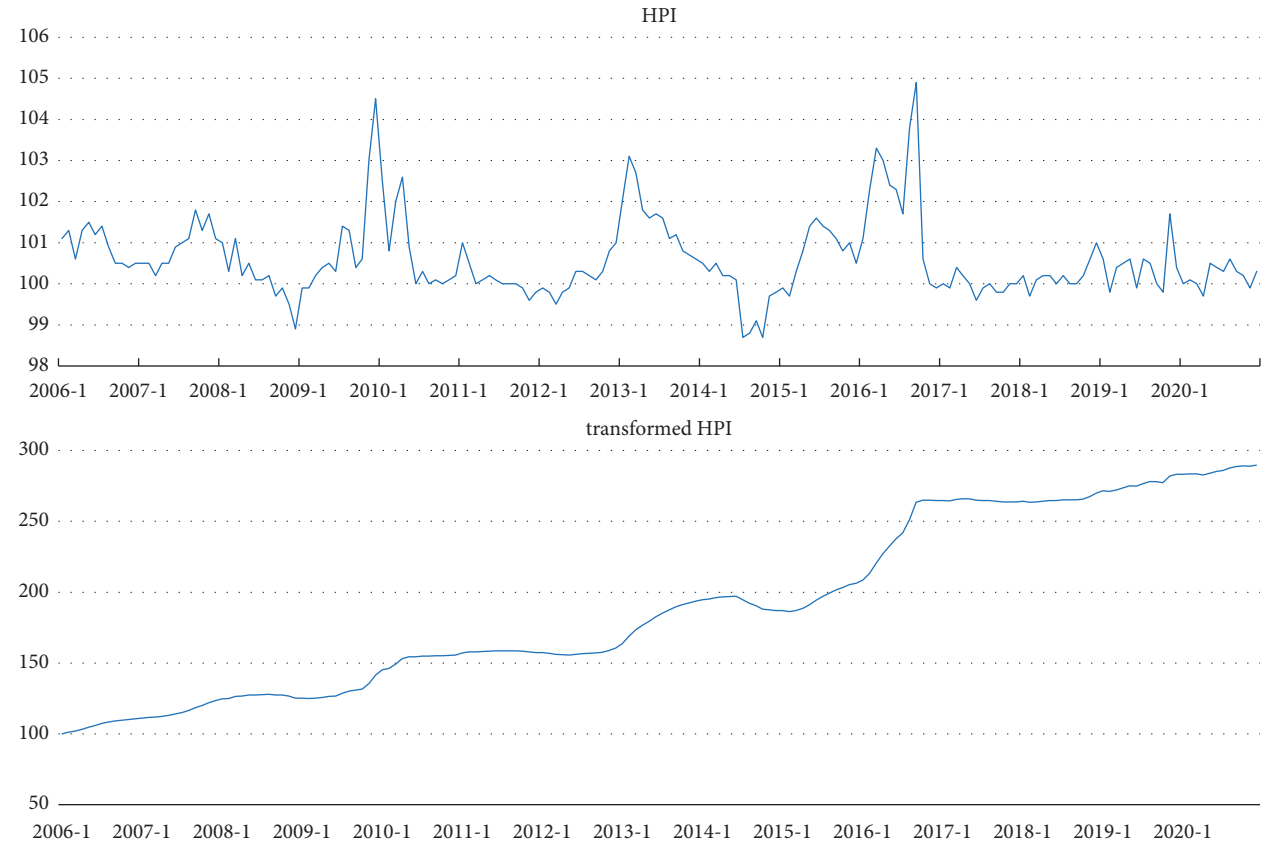


Figure 4: Comparison before and after data transformation.

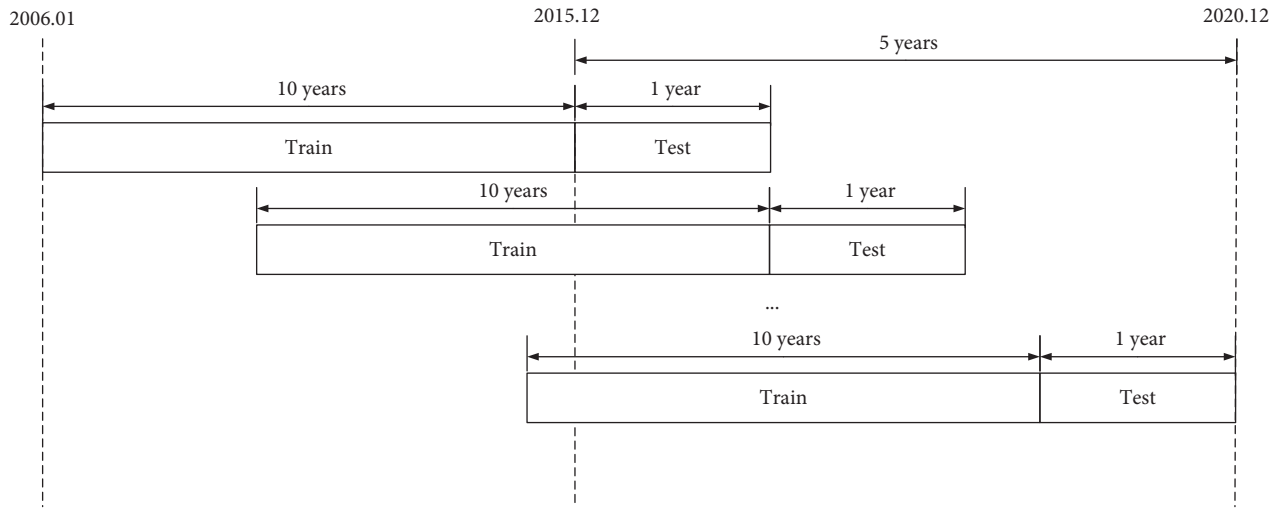


Figure 5: Segmented prediction process.

Table 4: Time interval division of data sets in different years.

Training set	Testing set
2006.1–2015.12	2016.1–2016.12
2007.1–2016.12	2017.1–2017.12
2008.1–2017.12	2018.1–2018.12
2009.1–2018.12	2019.1–2019.12
2010.1–2019.12	2020.1–2020.12

general idea of the DM test is as follows: assume that the prediction error $u_{i,\,t}$ is expressed as

$$u_{i,t} = \widehat{y}_{i,t} - y_t, \tag{18}$$

where $\widehat{y}_{i,t}$ is the predicted value and y_t is the true time series.

In the DM test, the null hypothesis that there is no significant difference in the predictive ability of the two

models is $H_0 : E(d_t) = 0$, where d_t is the relative loss function between the models and the expression is $dt = g(u_{1,t}) - g(u_{2,t})$ and $g(\cdot)$ is the loss function. The DM statistic is calculated as follows:

$$\mathrm{DM} = \frac{\overline{d}}{\sqrt{2\pi \widehat{f_d}(0)/T}}, \tag{19}$$

where $\overline{d} = 1/T \sum_{t=1}^{T} (g(u_{1,t}) - g(u_{2,t}))$ is the mean value of the loss difference, and $\widehat{f_d}(0)$ denotes the consistent estimate of f_d (0), which represents the spectral line density when the loss difference frequency is zero. In this paper, we use MSE and MAE as loss functions for DM tests, respectively.

3.4.4. Model Selection and Parameter Setting. All experiments presented in this paper are implemented based on python 3.7 software, and multiple algorithms are used as comparison models to measure the prediction performance of the bagging-WOA-SVR model in four sets of house price data. For single-model prediction, DecisionTreeRegressor (DTR), BPNN, and SVR are used to compare and analyze the prediction effect of the WOA-SVR model. Among the integrated algorithms, Random Forest (RF), Bagging-BPNN, and Bagging-SVR models are selected to compare and analyze the prediction effect of the bagging-WOA-SVR model, where the range of WOA algorithm parameters is set from 0.001 to 10. The parameters of each model are shown in Table 5.

4. Experiment and Discussion

4.1. Analysis of Model Prediction Effects. Tables 6 to 9 show the prediction effects of MSE and MAE for each model in the four-city dataset, while in order to further analyze the prediction effects of each model, the table shows the prediction index values for each year and the 5-year sum together, where the bolded font indicates the model with the best prediction effect in that year.

From the results in the table, it can be seen that, in the prediction index results of all years, the prediction results of the models after bagging integration are better than their corresponding single models, while the bagging-WOA-SVR model has the best prediction effect among the four data sets, which fully proves that the bagging integration strategy can effectively improve the model prediction accuracy. However, the bagging-WOA-SVR model does not always produce the best prediction results in each year's forecasting, but the model is able to achieve high accuracy in predicting the house price index. For example, in the prediction results of the Beijing dataset, although the MSE results of the bagging-BPNN model in the third and fifth years performed well, in the first and second years' predictions, the bagging-WOA-SVR improved by 74.25% and 84.31%, respectively, relative to the bagging-BPNN, which significantly outperformed the bagging-BPNN. Therefore, the bagging-WOA-SVR model has better stability and can achieve high accuracy forecasting of house price index data.

In the single-model prediction, the WOA-SVR model performed the best in the four data sets. For example, in the Beijing dataset, the MSE index results of the WOA-SVR model for all years improved by 40.96% relative to the SVR model, which fully demonstrates that the parameter settings have a large impact on the prediction effect of the SVR model, and the SVR model optimized by WOA can effectively improve the prediction accuracy. Meanwhile, compared with DTR and BPNN, the accuracy of MSE of the WOA-SVR model improved by 68.32% and 72%, respectively, which proved that, compared with the neural network and decision tree model, the SVR model after parameter search has a better prediction effect when performing house price index prediction.

The prediction curves of each model in the four data sets are shown in Figure 6.

4.2. DM Test. Table 10 shows the DM test results of the bagging-WOA-SVR model over the other models in the four data sets. In the DM test, when the p value is less than 0.05, it indicates that the null hypothesis is rejected and the predictive ability of the two models is significantly different. The boldface values in the table indicate that the p value is less than 0.05. In order to directly express the predictive abilities of the bagging-WOA-SVR model and other models, the predictive ability is analyzed using the coverage ratio based on the DM results. The expression of coverage is as follows: the number of DM results rejecting the null hypothesis/the total number of DM results. When the models have similar predictive ability, there are fewer DM test results with p values less than 0.05 and the coverage rate is less than 50%; when the models have significantly better predictive ability than the benchmark model, there are more DM test results with p values less than 0.05 and the coverage rate is greater than 50%.

In this way, further analysis of the DM test results in Table 10 reveals that the DM coverage of the bagging-WOA-SVR model is 100%, 85.71%, 92.86%, and 92.86% in the four data sets, respectively, whichproves that the bagging-WOA-SVR model significantly outperforms the proposed benchmark model in most cases; therefore, the proposed integrated model in this paper is statistically significant.

4.3. Feature Importance Analysis. To identify the most influential house price factors in making house price index forecasts for the four cities, the permutation_importance function in the sklearn package is used to perform a feature importance analysis on the WOA-SVR model that performs the best in single-model forecasting. Among them, since this paper uses a segmented forecasting method to forecast the 5-year house price index of the four cities and uses a rolling forecast for each stage to forecast the house price index of the future month with the characteristic data of the past quarter, the same characteristics of different months are numbered when performing the characteristic importance analysis. As an example of the house price index for the previous period, HPI_pr_1, HPI_ pr_2, and HPI_ pr_3 are used to represent the house price index from the first to the third month of the quarter.

Table 5: Model parameter settings.

	Beijing	Shanghai	Tianjin	Chongqing
Bagging-WOA-SVR	n_estimators = 13	n_estimators = 28	n_estimators = 10	n_estimators = 15
RF	n_estimators = 36, max_depth = 2, min_samples_leaf = 0.26	n_estimators = 30, max_depth = 2, min_samples_leaf = 0.3	n_estimators = 30, max_depth = 2, min_samples_leaf = 0.3	n_estimators = 18, max_depth = 2, min_samples_leaf = 0.32
Bagging-BPNN	n_estimators = 20	n_estimators = 18	n_estimators = 20	n_estimators = 17
Bagging-SVR	n_estimators = 36	n_estimators = 11	n_estimators = 10	n_estimators = 10
Population size of WOA	20	10	10	20
Number of WOA iterations	100	50	100	100
SVR	Defult	Defult	Defult	Defult
BPNN	Number of neurons in the hidden layer = 13; batch_size = 256; epoch = 50	Number of neurons in the hidden layer = 13; batch_size = 256; epoch = 50	Number of neurons in the hidden layer = 13; batch_size = 256; epoch = 50	Number of neurons in the hidden layer = 13; batch_size = 256; epoch = 50
DTR	Defult	Defult	Defult	Defult

Table 6: Forecasting accuracy of house price index in Beijing.

	MSE						MAE					
	Year 1	Year2	Year 3	Year 4	Year5	All years	Year 1	Year 2	Year 3	Year 4	Year 5	All years
Bagging-WOA-SVR	**1.2358**	0.4084	1.5065	**1.3260**	1.3260	**1.1696**	**1.1117**	0.4824	0.9128	1.2274	1.0243	**0.9517**
Bagging-SVR	4.2971	1.5849	2.5588	2.9732	2.3655	2.7559	2.0129	1.0732	1.4988	1.6615	1.4853	1.5463
RF	3.3657	**0.2774**	0.7305	5.4396	2.0001	2.3627	1.7715	**0.2557**	**0.4279**	1.9242	1.0438	1.0846
Bagging-BPNN	4.7993	2.6036	**0.3903**	1.5703	**0.7512**	2.0229	2.1325	1.4722	0.5101	**1.1968**	**0.7508**	1.2125
WOA-SVR	3.4919	1.8743	1.4930	2.3628	3.0009	1.8186	1.8348	1.3640	1.1520	1.5311	1.6922	1.2321
SVR	8.6311	1.1002	1.3480	2.4745	1.8485	3.0805	2.8398	0.9686	1.0412	1.5234	1.3178	1.5382
DTR	12.4391	2.9000	2.3459	5.8592	5.1573	5.7403	2.9587	1.4080	0.8401	1.4989	1.6650	1.6741
BPNN	11.0144	3.0294	2.6407	9.3291	6.4641	6.4960	2.7498	1.44217	1.2545	2.6853	1.9111	2.0087

Table 7: Forecasting accuracy of house price index in Shanghai.

	MSE						MAE					
	Year 1	Year 2	Year 3	Year 4	Year 5	All years	Year 1	Year 2	Year 3	Year 4	Year 5	All years
Bagging-WOA-SVR	0.8155	0.4948	0.5696	0.6774	**0.9878**	**0.7090**	0.9030	0.5956	0.6389	0.6511	**0.8746**	**0.7326**
Bagging-SVR	3.1095	1.1233	1.8195	1.8327	2.0666	1.9157	1.7489	0.9722	1.2976	1.3232	1.3974	1.3150
RFR	1.6213	0.3555	0.9236	2.8809	2.6231	1.6809	1.2288	**0.2552**	0.5455	1.3989	1.3721	0.9601
Bagging-BPNN	**0.7090**	7.2293	0.6733	**0.4677**	2.2952	2.2749	**0.8077**	2.6741	0.7184	**0.5878**	1.3076	1.2191
WOA-SVR	0.8541	0.4224	0.5157	0.7794	1.1578	0.7459	0.9241	0.5189	0.6366	0.6787	1.0551	0.7627
SVR	3.8000	1.0251	1.5441	1.6850	1.9462	2.0001	1.9267	0.8224	1.1579	1.2280	1.3375	1.2945
DTR	8.9016	**0.1747**	**0.3583**	0.5111	4.0907	2.8073	2.7198	0.2711	**0.4509**	0.5980	1.4023	1.0884
BPNN	7.5682	2.9868	1.5620	5.9379	9.4485	5.5304	2.2624	1.3066	1.1304	1.9871	2.5365	1.8489

Table 8: Forecasting accuracy of house price index in Tianjin.

	MSE						MAE					
	Year 1	Year 2	Year 3	Year 4	Year 5	All years	Year 1	Year 2	Year 3	Year 4	Year 5	All years
Bagging-WOA-SVR	**0.1304**	0.4395	0.3722	**0.3187**	**0.1812**	**0.2884**	**0.3611**	0.5302	0.6101	0.5644	0.4253	**0.4982**
Bagging-SVR	0.4365	0.9004	0.9893	0.7963	1.0836	0.8412	0.6256	0.7859	0.9259	0.7825	0.9866	0.8213
RF	0.8113	**0.1583**	0.5530	0.5904	0.6158	0.5458	0.9007	**0.2413**	0.5156	0.5446	0.6098	0.5624
Bagging-BPNN	0.2971	2.4291	**0.2351**	1.1269	0.2012	0.8632	0.4853	1.5397	**0.4261**	1.0194	**0.3738**	0.7722
WOA-SVR	0.3791	0.4986	0.3238	0.6699	0.2397	0.4222	0.5666	0.7026	0.5687	0.7594	0.4893	0.6173
SVR	0.4245	0.7765	1.1729	0.9312	0.9986	0.8607	0.5729	0.7404	0.9689	0.8462	0.9119	0.8081
DTR	1.1372	1.4663	2.1395	0.4295	0.3560	1.1057	1.0221	0.8026	0.7709	**0.4728**	0.4592	0.7055
BPNN	3.8443	1.3129	1.2982	1.7157	1.9736	2.0251	1.7255	0.8984	0.7463	1.1382	1.1482	1.1294

Table 9: Forecasting accuracy of house price index in Chongqing.

	MSE						MAE					
	Year 1	Year 2	Year 3	Year 4	Year 5	All years	Year 1	Year 2	Year 3	Year 4	Year 5	All years
Bagging-WOA-SVR	**0.0771**	**0.1336**	0.2134	0.3119	0.3681	**0.2208**	**0.2711**	**0.3649**	0.4619	0.5581	0.6066	**0.4525**
Bagging-SVR	0.4056	0.2022	0.5983	0.6238	0.4174	0.4494	0.4956	0.4211	0.7073	0.7399	0.6192	0.5966
RF	0.1838	0.2332	0.3137	0.4819	0.4047	0.3235	0.4223	0.4772	0.5601	0.6870	0.6235	0.5540
Bagging-BPNN	0.3417	1.0878	**0.1557**	**0.1355**	**0.2613**	0.3964	0.5411	0.9642	**0.3275**	**0.3205**	**0.4269**	0.5160
WOA-SVR	0.2207	0.2629	0.1878	0.4015	0.4294	0.3005	0.4588	0.5114	0.4334	0.6335	0.5881	0.5250
SVR	0.3850	0.1982	0.8869	0.6998	0.4362	0.5212	0.5014	0.3996	0.8629	0.7805	0.6174	0.6324
DTR	2.2237	0.6949	0.7893	1.2810	2.5617	1.5101	0.8314	0.6472	0.7505	1.0542	1.2425	0.9052
BPNN	1.3845	0.9405	1.4364	2.1968	2.0749	1.6110	0.8765	0.7458	1.0868	1.1973	1.0842	0.9981

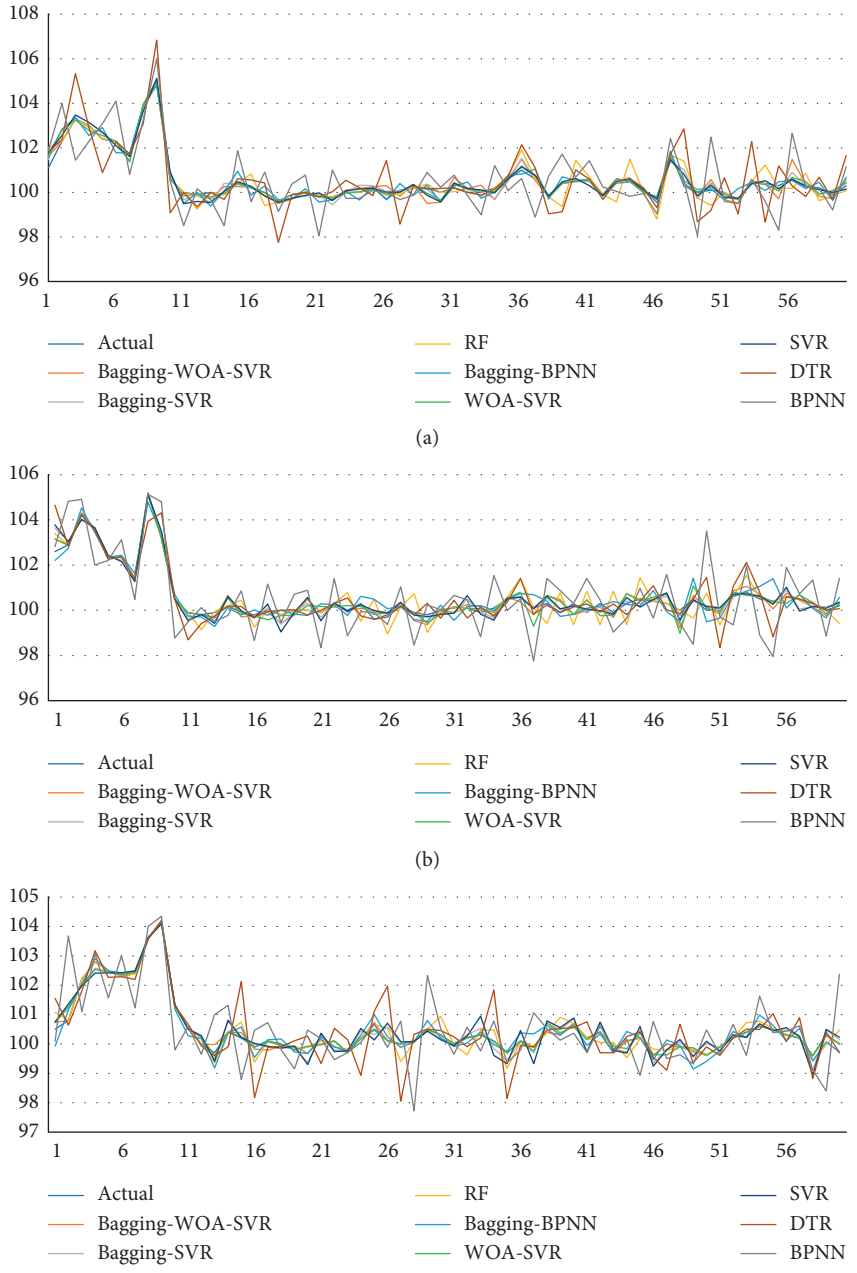
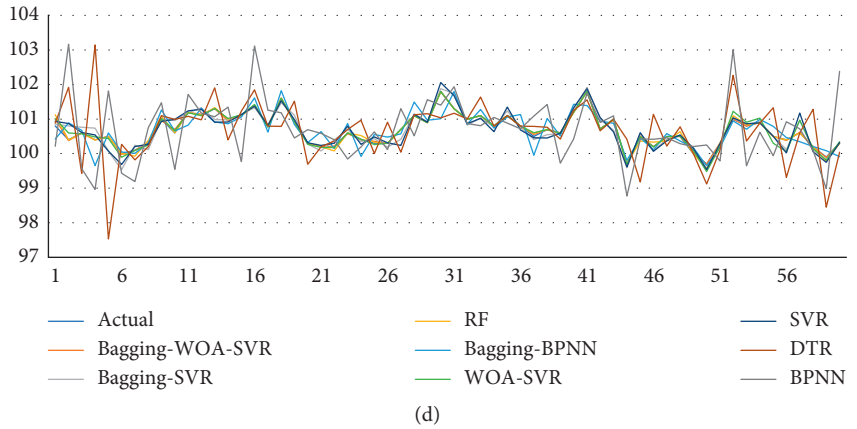


Figure 6: Continued.



(d)

Figure 6: Fitting curves for four cities. (a) Fitting curve of house price index in Beijing. (b) Fitting curve of house price index in Shanghai. (c) Fitting curve of house price index in Tianjin. (d) Fitting curve of house price index in Chongqing.

Table 10: DM test results for bagging-WOA-SVR model and comparison model.

	B-SVR	RF	B-BPNN	WOA-SVR	SVR	DTR	BPNN
				Beijing			
MSE	**5.4556e−07**	**0.0002**	**0.0072**	**1.6789e−07**	**0.0071**	**0.0063**	**0.0013**
MAE	**3.6952e-11**	**0.0031**	**0.0093**	**1.9554e−11**	**8.4994e−05**	**0.0008**	**2.2731e−05**
				Shanghai			
MSE	**1.1166e−08**	**1.3173e−06**	**0.0004**	0.8486	**2.8626e−06**	**0.0009**	**0.0008**
MAE	**1.7460e−14**	**4.0219e−05**	**0.0002**	0.6044	**3.1764e−10**	**0.0014**	**7.0947e−06**
				Tianjin			
MSE	**2.3444e−07**	**9.0037e−05**	**0.0001**	**0.0944**	**0.0007**	0.0623	**0.0093**
MAE	**1.6335e−09**	**0.0030**	**2.2904e-05**	**0.0041**	**4.7335e−07**	**0.0172**	**2.5020e−05**
				Chongqing			
MSE	**0.0030**	**2.2542e−08**	**0.0037**	**0.0122**	**0.0007**	0.0996	**0.0154**
MAE	**7.1201e−05**	**1.5571e−10**	**0.0157**	**0.0004**	**1.6134e−05**	**0.0019**	**0.0001**

Table 11: Feature importance analysis of Beijing dataset.

2016.1–2016.12		2017.1–2017.12		2018.1–2018.12		2019.1–2019.12		2020.1–2020.12	
Factor	Score	Factor	Score	Factor	Score	Factor	Score	Factor	Score
HPI_pr_3	0.242815	HPI_pr_3	0.43138	HPI_pr_3	0.549026	ERateIndex_1	0.004201	HPI_pr_3	0.323126
HPI_pr_2	0.071025	ERateIndex _2	0.04204	HPI_pr_1	0.076841	ERateIndex_2	0.003756	HPI_pr_2	0.04894
ERateIndex _3	0.064863	HPI_pr_2	0.032962	ERateIndex_2	0.054605	HPI_pr_3	0.003746	HPI_pr_1	0.032421
Investmentcom _3	0.058553	M2 _2	0.031422	ERateIndex_3	0.046494	M2 _1	0.003728	ERateIndex_2	0.027879
M2 _2	0.051563	ERateIndex_3	0.030907	M2 _2	0.038579	CPI_2	0.003727	M2 _2	0.02218

Table 12: Feature importance analysis of Shanghai dataset.

2016.1–2016.12		2017.1–2017.12		2018.1–2018.12		2019.1–2019.12		2020.1–2020.12	
Factor	Score	Factor	Score	Factor	Score	Factor	Score	Factor	Score
HPI_pr_3	0.003742	HPI_pr_3	0.5881	HPI_pr_3	0.322272	HPI_pr_3	0.547713	HPI_pr_3	0.29674
HPI_pr_2	0.002568	HPI_pr_2	0.076124	HPI_pr_2	0.08023	HPI_pr_1	0.094856	HPI_pr_2	0.096047
HPI_pr_1	0.001568	M2_1	0.037817	HPI_pr_1	0.053294	HPI_pr_2	0.084988	HPI_pr_1	0.057768
M2 _1	0.000676	HPI_pr_1	0.037011	CPI_3	0.016268	M2_1	0.066755	ERateIndex _3	0.016412
Population _1	0.000393	ERateIndex _3	0.014378	M2_1	0.011593	M2_2	0.058954	CPI_2	0.008884

The characteristics of the WOA-SVR model that rank the top 5 in importance for each year when forecasting the five-year house price index for the four cities are listed in Table 11 to 14. Further analysis of the results shows the following.

(1) The house price index as a typical time series has characteristics such as stochasticity and trend, and information on the house price index in previous periods is usually the best source of information when forecasting the current period

Table 13: Feature importance analysis of Tianjin dataset.

2016.1–2016.12		2017.1–2017.12		2018.1–2018.12		2019.1–2019.12		2020.1–2020.12	
Factor	Score	Factor	Score	Factor	Score	Factor	Score	Factor	Score
ERateIndex _2	0.119247	HPI_pr_3	0.073616	HPI_pr_3	0.014421	HPI_pr_3	0.341244	HPI_pr_3	0.028759
ERateIndex _3	0.109587	HPI_pr_2	0.035202	HPI_pr_2	0.007882	HPI_pr_2	0.061115	HPI_pr_2	0.020244
M2 _3	0.096798	ERateIndex _3	0.022461	HPI_pr_1	0.005814	ERateIndex _3	0.044221	ERateIndex _3	0.019735
CPI_1	0.089855	HPI_pr_1	0.018916	ERateIndex _2	0.003456	CPI_3	0.043533	ERateIndex _2	0.019326
CPI_2	0.059374	ERateIndex _2	0.008232	ERateIndex _3	0.003325	ERateIndex_2	0.035806	ERateIndex _1	0.018893

Table 14: Feature importance analysis of Chongqing dataset.

2016.1–2016.12		2017.1–2017.12		2018.1–2018.12		2019.1–2019.12		2020.1–2020.12	
Factor	Score	Factor	Score	Factor	Score	Factor	Score	Factor	Score
HPI_pr_3	0.068482	HPI_pr_3	0.045737	HPI_pr_3	0.000103	HPI_pr_3	0.011434	HPI_pr_3	0.473192
HPI_pr _2	0.027638	HPI_pr _2	0.017391	HPI_pr _2	0.000069	HPI_pr _2	0.008818	HPI_pr _2	0.185356
CPI _2	0.022059	ERateIndex _3	0.012543	HPI_pr _2	0.000046	HPI_pr _2	0.006055	ERateIndex _3	0.047204
ERateIndex _3	0.011095	CPI_1	0.010967	ERateIndex _3	0.000020	ERateIndex _3	0.00249	ERateIndex _1	0.033405
CPI_1	0.009163	CPI _2	0.010454	ERateIndex _1	0.000011	ERateIndex _2	0.001032	M2 _2	0.013421

index. Thus, in the characteristic ranking results for all four cities, the previous period house price index appears in the top 5 when forecasting for each year, and the results show that HPI_pr_3, the previous period house price index closest to the forecasting month, is in the top of the characteristic ranking for all but Beijing 2019 and Tianjin 2016. (2) The exchange rate index is found in the top five characteristics of Beijing, Tianjin, and Chongqing, to varying degrees. Specifically, the influx of international capital will continue to increase liquidity in the financial markets, resulting in lower interest rates and lower purchase costs, to promote housing prices. (3) The results of ranking the importance of characteristics in Beijing and Shanghai show that money supply is also an important characteristic. The impact of money supply on housing prices is mainly manifested in the impact on interest rates. When the money supply increases, it leads to a decline in the market equilibrium interest rate, which is indirectly transmitted to the real estate industry. At the same time, the money supply will also have a direct impact on the real estate market. When the money supply changes, it will directly cause the price of raw materials for housing production to change, which will have an impact on housing prices. (4) In the feature ranking results of the four sets of data, in addition to the above three factors, CPI also plays an important role in the prediction of some years. CPI reflects the changes in the prices of consumer goods over a period of time, and usually, its rate of change can reflect the level of inflation in the region.

5. Conclusions

This paper forecasts the house price indices based on four data sets in Beijing, Shanghai, Tianjin, and Chongqing. Firstly, the literature analysis method and gray correlation analysis method are used to establish the forecasting index system, and the segmented forecasting method is used to forecast the house price index in the coming year with the data information of the past ten years, and then, the bagging integrated WOA-SVR model is used to analyze the house price indices of the four cities in experimental simulations. Conclusions can be made as follows:

(i) Compared with other machine learning models, the bagging-WOA-SVR model proposed in this paper can effectively improve the prediction accuracy of the house price index. By analyzing the 5-year forecasting results of all models, the integrated model proposed in this paper has better stability and can achieve high-precision forecasting of house price indexes

(ii) The DM test with the comparison model on the four data sets found that the integrated model proposed in this paper has a higher coverage rate and significantly outperforms the comparison model in most cases, which is statistically significant

(iii) By analyzing the feature importance of the four cities, it is found that the house price index of the previous period closest to the forecast month, exchange rate index, M2, and CPI are important influencing factors affecting the forecast of the house price index

Data Availability

The indicator data in this study are published on the website of the National Bureau of Statistics, except for the exchange rate index, which is obtained from the wind database.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (Grant no. 81973791).

References

[1] L. L. Zhao, J. Mbachu, and Z. S. Liu, "Exploring the trend of New Zealand housing prices to support sustainable development," *Sustainability*, vol. 11, no. 9, 2019.

[2] P. G. Hou and Z. Q. Qiao, "Research on house price forecasting based on wavelet analysis and ARMA model," *Statistics & Decisions*, vol. 411, no. 15, pp. 20–23, 2014.
[3] Q. Y. Wang, "Analysis of fluctuation characteristics and trend prediction of house price and CPI in Beijing--analysis of GARCH family model based on cointegration relationship," *Price:Theory & Practice*, vol. 325, no. 7, pp. 57-58, 2011.
[4] Z. D. Li, J. H. Lin, and M. J. Wang, "High-dimensional statistics in big data era: development and application of sparse modeling," *Statistical Research*, vol. 32, no. 10, pp. 3–11, 2015.
[5] H. Y. Liao, J. P. Zeng, and C. R. Wu, "A model for online forum traffic prediction integrated with multiple models," *Computer Engineering*, vol. 46, no. 12, pp. 62–66, 2020.
[6] Q. Zhao, W. J. Xu, Y. C. Ji, G. Liu, and W. Zhang, "Application of machine learning to financial asset price forecasting and allocation: a literature review," *Chinese Journal of Management*, vol. 17, no. 11, pp. 1716–1218, 2020.
[7] S. Y. Liu, J. D. Huang, L. Q. Xu et al., "Combined model for prediction of air temperature in poultry house for lion-head goose breeding based on PCA-SVR-ARMA," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 36, no. 11, pp. 225–233, 2020.
[8] Y.-F. Ye, Y. H. Shao, Y.-H. Shao, and C.-N. Li, "WaveletLp-norm support vector regression with feature selection," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 19, no. 3, pp. 407–416, 2015.
[9] X. Han and L. Clemmensen, "On weighted support vector regression," *Quality and Reliability Engineering International*, vol. 30, no. 6, pp. 891–903, 2014.
[10] Q. Dong, N. N. Sun, and W. Li, "Real estate price prediction based on web search data," *Statistical Research*, vol. 31, no. 10, pp. 81–88, 2014.
[11] P.-F. Pai and W.-C. Hong, "Forecasting regional electricity load based on recurrent support vector machines with genetic algorithms," *Electric Power Systems Research*, vol. 74, no. 3, pp. 417–425, 2005.
[12] R. Liu and R. Hong, "Housing price forecasting based on least squares support vector regression with particle swarm optimization," *Journal of Convergence Information Technology*, vol. 6, no. 11, pp. 325–333, 2011.
[13] X. B. Tang, R. Zhang, and L. X. Liu, "Research on forecast of second-hand house price in beijing based on SVR model of bat algorithm," *Statistical Research*, vol. 35, no. 11, pp. 71–81, 2018.
[14] S. H. Choi and H. Jin, "An ensemble learner-based bagging model using past output data for photovoltaic forecasting," *Energies*, vol. 13, no. 6, 2020.
[15] M. Yang, J. Zhang, H. Lu, and J. Jin, "Regularized ELM bagging model for tropical cyclone tracks prediction in South China Sea," *Cognitive Systems Research*, vol. 65, pp. 50–59, 2021.
[16] S. Jung, J. Moon, S. Park, S. Rho, S. W. Baik, and E. Hwang, "Bagging ensemble of multilayer perceptrons for missing electricity consumption data imputation," *Sensors (Basel, Switzerland)*, vol. 20, no. 6, 2020.
[17] N. V. Vladimir, "The nature of statistical learning theory," *Technometrics*, vol. 38, no. 4, p. 400, 1996.
[18] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016.
[19] T. Gu, G. L. Xu, W. L. Li, J. Li, Z. Wang, and J. Luo, "Intelligent house price evaluation model based on ensemble LightGBM and Bayesian optimization strategy," *Journal of Computer Applications*, vol. 40, no. 9, pp. 2762–2767, 2020.
[20] F. F. Sun, "A brief discussion of the gray correlation analysis method and its application," *Science & Technology Information*, vol. 337, no. 17, pp. 880–882, 2010.
[21] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PLoS One*, vol. 12, no. 7, Article ID e0180944, 2017.