Prediction Of Vegetable Pricing and Replenishment Based on ARIMA Model

Dewang Meng *, Xingyu Jiang, Zhongjun Xie

Department of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China, 255000

* Corresponding Author Email: meng18732839566 @163.com

Abstract. Vegetables are indispensable consumer goods in the daily diet of the majority of consumers. Reasonable pricing and replenishment strategies of vegetables are conducive to maintaining the interests of agricultural producers and promoting the stable development of the market. The main contents of this paper are two points: one is to visualize the pre-processed data and establish the distribution map of vegetable sales in recent years; Secondly, the ARIMA time series prediction model was established by linear fitting of the data, and the feasibility test of the prediction results was carried out to obtain the total daily replenishment quantity and price of each vegetable category in a period of time. Under the background of the new era with low-carbon green as the mainstream, this research has broad development space and prospect. The pricing and replenishment strategies formulated in this study can reduce the waste of vegetables in the whole consumption supply chain, maintain the purchasing power and living standard of consumers, and meet the market demand. Marketing techniques that promote healthy foods in supermarkets are important to encourage healthy eating at the population level. The content of this study can provide ideas for the optimization of vegetable supply chain, further use of new technology, reduce waste, improve efficiency, and ensure the quality and supply stability of vegetables. This study is also helpful to analyze consumer behavior and understand consumers' reactions to vegetable price fluctuations and changes in purchasing behavior in order to better predict market trends and adjust production and supply strategies.

Keywords: ARIMA Model, Linear Regression, Correlation Analysis.

1. Introduction

In recent years, with the growth of population and the acceleration of urbanization, people's demand for food such as vegetables is also increasing. Faced with the increasingly large vegetable production and sales groups and vegetable consumption groups at home and abroad, vegetable price has become a key hub in vegetable supply and demand. However, in recent years, China's vegetable prices have fluctuated wildly, which not only distorts the allocation of market resources, but also damages the interests of vegetable producers, consumers and all parties in the whole circulation process, and even affects social stability [3]. The supply level of vegetables and its price level have an important impact on agricultural economic benefits and the health and stability of the national economy. At the same time, the level and stability of vegetable prices have an important impact on the consumption level and quality of life of urban and rural residents [4]. The price index of agricultural products is an important economic index that reflects the direction, change trend and fluctuation degree of agricultural products market in a certain period [5]. This study has far-reaching significance for theoretical research: the ARIMA time series prediction model can be used to further verify and improve the application effect of this model in agricultural products market, provide a new application scenario for time series analysis, and provide a reference for other researchers to explore time series problems in different fields. It also provides a new research perspective and method for different fields and can use the principles of economics and statistics to explore the competitive factors in the vegetable market combined with the results of this study. This research has far-reaching significance to practical life: First, it can effectively balance the market supply and demand relationship, reduce the waste of vegetables, and ensure the stability of the market. Second, to help consumers make more informed purchasing decisions, thereby protecting their interests; Finally, it

provides reasonable data support for government departments, which is helpful to formulate more reasonable and effective agricultural policies. To sum up, the prediction of vegetable price and supply based on ARIMA model has important theoretical and practical significance.

2. The basic fundamental of model

2.1. The structure of Curve fitting model

Curve fitting is a method commonly used in statistics. It refers to finding the best function curve or model to describe the trend and relationship of data by fitting data to a given set of data, so as to predict the future or perform data analysis. As the most commonly used regression analysis method in statistics, linear regression has been fully recognized and widely used [6]. Therefore, a linear regression model will be used for fitting in this study.

Because nonlinear regression analysis can be transformed into linear regression problem by the transformation of function equation, linear regression analysis is the most basic and more widely used method in practical application [7]. It assumes that there is a linear relationship between the independent variable and the dependent variable, and that the error term follows a normal distribution. Linear regression can be used to predict and establish relationships between dependent and independent variables. Its important parameters are equation slope p1 and intercept p2, fitting square variance SSE, fitting standard deviation RMSE, fitting degree R^2, adjustment R^2, etc. Here is an explanation of SEE, R^2, adjusting R^2:

(1) SSE: The closer SSE is to 0, the better model selection and fitting, and the more successful data prediction.

(2)R^2: the degree of fitting of the curve regression, in general, the closer to 1 the better the effect.

(3) Adjusted R^2: It will not increase with the increase of the number of independent variables. In multiple linear regression analysis, it will be more accurate to use adjusted R^2 to evaluate the model fitting effect.

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{1}$$

$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{2}$$

$$R = \frac{\sum (x - \bar{x})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})}} \tag{3}$$

$$R^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-k-1)} \tag{4}$$

2.2. The structure of ARIMA time series prediction model

ARIMA Model is Autoregressive Integrated Moving Average Model, in which ARIMA (p, d, q) is called differential autoregressive moving average model, and is one of the most widely used time series modeling methods at present [8]. Time series is a series obtained by arranging a set of data according to a certain time interval. It is the ultimate purpose of time series analysis to analyze and find out the change rule and the internal quantitative relationship and use this information to predict future data based on historical data [9]. Differential integrated moving average autoregressive model (ARIMA model) can predict the future development according to the existing historical data and is an organic combination of autoregressive model (AR model) and moving average model (MA model) [10]. ARIMA model is mainly composed of three parts, namely, autoregressive model (AR),

differential process (I) and moving average model (MA). Here's an explanation of the three components:

(1) AR: refers to the past value used to predict the next value. The AR term is defined by the parameter p in ARIMA. The P-value is determined by the PACF diagram.

(2)I: specifies the number of differential operations performed on the sequence, the purpose of differential operations on the data is to keep it stable. ADF can be used to determine whether a sequence is stationary or not and to help identify D-values.

(3) MA: defines the number of past prediction errors when predicting future values. The parameter q in ARIMA represents the MA term. ACF plots are used to identify the correct q values.

$$X_t = c + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \ldots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} \tag{5}$$

$$X_t = w_t, w_t(0, \sigma^2) \tag{6}$$

$$X_t = X_{t-1} + w_t, X_t(0, t\sigma^2) \tag{7}$$

$$\Delta X = X_t - X_{t-1} = w_t, U_t N(0, \sigma^2) \tag{8}$$

$$\Delta X = X_t - X_{t-1} = w_t, U_t N(0, \sigma^2) \tag{9}$$

$$PACF(k) = \frac{E(Z_t - EZ_t)\ (Z_{t-k} - EZ_{t-k})}{\sqrt{E(Z_t - EZ_t)^2}\sqrt{E(Z_{t-k} - EZ_{t-k})^2}} = \frac{\mathrm{cov}[(Z_t - \bar{Z}_t), (Z_{t-k} - \bar{Z}_{t-k})]}{\sqrt{\mathrm{var}(Z_{t-k} - \bar{Z}_{t-k})}\sqrt{\mathrm{var}(Z_t - \bar{Z}_t)}} \tag{10}$$

3. Results

3.1. The establishment of simulation model

3.1.1 Data collection and visualization processing

The data collected in this study are:

Commodity information of 6 vegetable categories and 252 single items in a supermarket.

The replenishment volume and pricing of each item of 6 vegetable categories in a supermarket from July 2020 to July 2023 (excluding special treatment items)

the wholesale price of each item of 6 vegetable categories in a supermarket from July 2020 to July 2023

And before the study, we used the normal distribution 3σ principle to clear abnormal data to reduce the error of experimental data.

According to the known data, Microsoft Excel software was used to import the sales data of vegetable categories and each single product, and visualization processing was carried out to obtain the total sales diagram of each vegetable category in different time periods and the total sales diagram of each category in a single day within 24 hours, as shown in Figure 1:

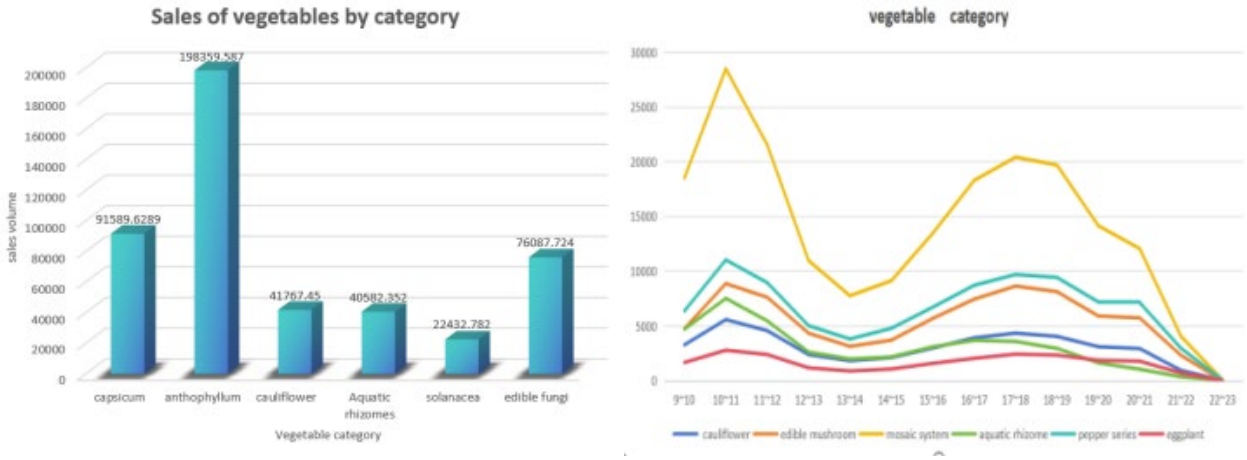


Figure 1. Chart of total sales and daily sales

3.1.2 Linear fitting of data

According to the collected data, the independent variable of the linear fitting model is first defined as the unit price of sales, and the dependent variable is the sales volume. Then Matlab software is used to calculate the optimal slope m and intercept b to minimize the residual sum of squares between the observed data points and the predicted value of the model. According to the calculated slope and intercept, a fitting line is drawn and displayed on the chart together with the original data points to achieve a visual fitting effect. The statistical index R^2 is used to measure how well the fitting line fits the data, and the residual is checked to ensure that the model conforms to the assumptions of the data. Finally, within the permissible range of random error, the research conclusion is drawn.

Since there are too many charts and other kinds of vegetables all show similar fitting degree, only the linear fitting graphs of cauliflower and Mosaic are given here, as shown in Figure 2:

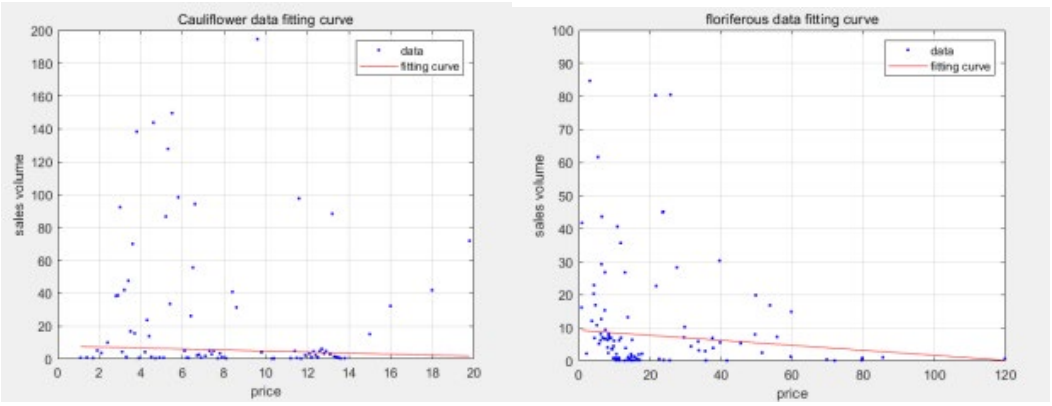


Figure 2. Linear fitting graph

3.1.3 Establishment of ARIMA time series prediction model

Because there are too many charts, and the prediction degree of the six kinds of vegetables is similar, only the time series forecast of the sales volume and pricing of flowers and leaves is given here.

First of all, the data is standardized, and the ADF test is carried out on the series plot to observe whether the series is stable (generally unstable). For the non-stationary time series, D-order difference is carried out first to transform it into a stationary time series and determine the D-value.

The ADF inspection table is shown in Table 1:

Table 1. The ADF inspection table

variable	Difference order	t	P	AIC	critical value		
					1%	5%	10%
Replenishment volume	0	-2.809	0.057*	-3439.087	-3.437	-2.864	-2.568
	1	-11.752	0.000***	-3429.797	-3.437	-2.864	-2.568
	2	-15.062	0.000***	-3330.699	-3.437	-2.864	-2.568
pricing	0	-3.095	0.027**	-2337.311	-3.437	-2.864	-2.568
	1	-8.473	0.000***	-2327.26	-3.437	-2.864	-2.568
	2	-14.8	0.000***	-2255.458	-3.437	-2.864	-2.568

Note: * * *, * * and * represent the significance levels of 1%, 5% and 10% respectively.

Secondly, look at the data comparison chart before and after the difference to judge whether it is stable (the fluctuation amplitude is not large), and at the same time, make partial (autocorrelation analysis) on the time series, and estimate its P and Q values according to the truncation, in which: the P value can be roughly judged from the maximum lag point of the partial autocorrelation coefficient (PACF) chart, and the Q value can be roughly judged from the maximum lag point of the autocorrelation coefficient (ACF) chart.

The best difference sequence diagram for Quantity of flower and leaf vegetable replenishment, final difference data autocorrelation diagram (ACF) is shown in Figure 3:

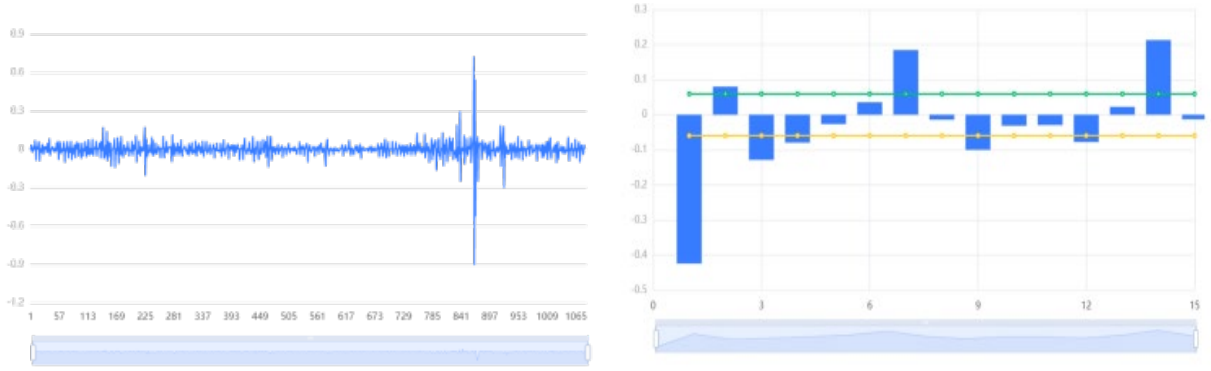


Figure 3. Optimal difference sequence diagram and ACF

The best difference sequence diagram for Mosaic pricing, final difference data autocorrelation diagram (ACF) is shown in Figure 4:

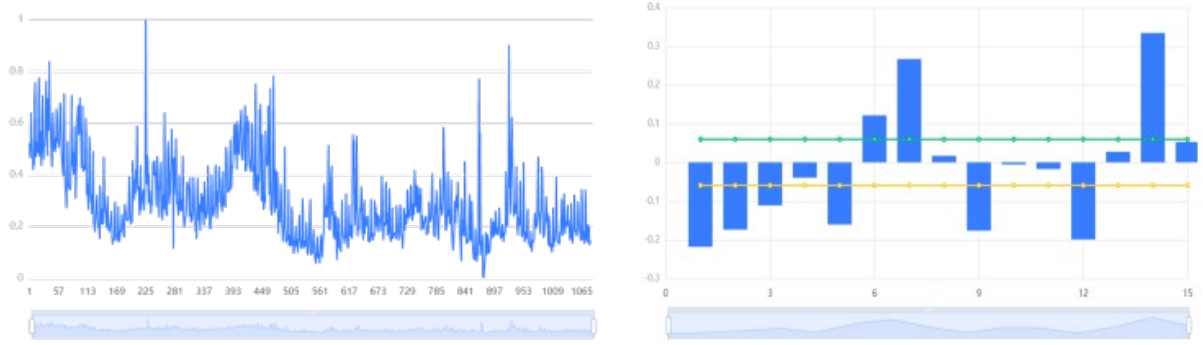


Figure 4. Optimal difference sequence diagram and ACF

Third, fit the ARIMA model. The ARIMA model requires the model to have pure randomness, that is, the model residual is white noise. Check the model test table and test the model white noise according to the P-value of the Q statistic (P>0.05) or analyze the information criteria AIC and BIC values.

The ARIMA model test table of Mosaic replenishment quantity and total pricing is shown in Table 2:

Table 2. The ARIMA model test table

Replenishment volume			pricing		
Item	sign	value	Item	sign	value
Sample size	Df Residuals	1076	Sample size	Df Residuals	1079
	N	1085		N	1085
Q statistic	Q6	0.068(0.795)	Q statistic	Q6	0.047(0.829)
	Q12	7.78(0.255)		Q12	35.97(0.000***)
Information criterion	AIC	-3420.214	Information criterion	AIC	-2123.308
	BIC	-3375.319		BIC	-2093.378
Goodness of fit	R^2	0.497	Goodness of fit	R^2	0.633

Finally, according to the model parameter table and combined with the time series analysis chart, the order result of backward prediction is obtained, that is, the predicted value of the total price of vegetables and the quantity of replenishment in this study.

The forecast figure of total pricing and replenishment of Mosaic and leaf vegetables is shown in Figure 5:

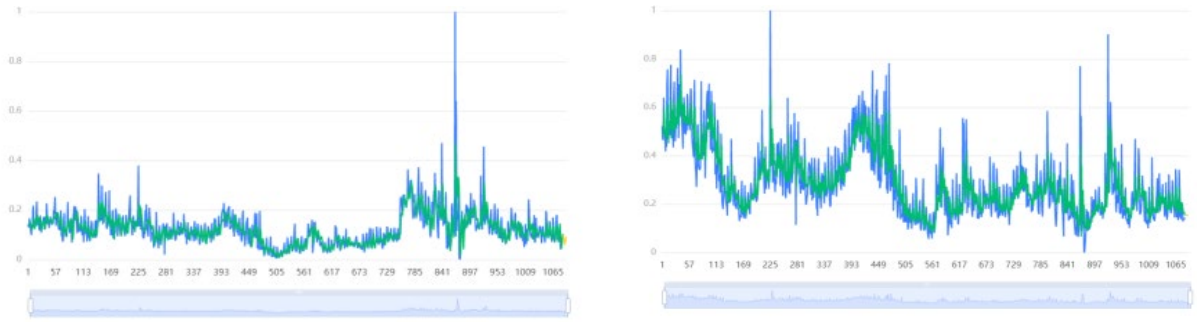


Figure 5. The forecast figure of total pricing and replenishment

The time series prediction table of Mosaic category pricing and replenishment volume in the next seven days is shown in Table 3:

Table 3. Predicted value.

Order (time)	Replenishment volume	Total pricing
1	125.5349	1019.618
2	108.9192	1004.906
3	89.1905	1004.266
4	75.9761	984.437
5	81.6293	981.878
6	99.90276	979.959
7	116.8373	977.400

Predicted values of 6 kinds of vegetables as shown in Table 4:

Table 4. Predicted values of 6 kinds of vegetables

	Total replenishment(kg)						Total pricing(yuan/kg)					
days	Cauliflower	Mosaic	chili	Edible fungi	Eggplant	Aquatic root	Cauliflower	Mosaic	chili	Edible fungi	Eggplant	Aquatic root vegetables
1	27.86	125.5	74.15	43.84	21.10	19.64	602.6	1019	923.6	843.7	369.3	720.8
2	22.03	108.9	71.65	46.93	21.07	19.31	554.6	1004	926.6	764.4	369.3	654.1
3	20.21	89.19	69.57	48.82	21.06	19.86	535.1	1004	927.3	734.9	369.3	661.9
4	19.63	75.98	68.60	48.83	21.05	20.08	527.0	984.1	926.9	730.3	369.3	651.6
5	19.43	81.63	71.98	48.55	21.03	20.30	523.4	981.8	925.7	731.1	369.3	647.6
6	19.35	99.90	73.64	48.28	21.02	20.51	521.5	979.9	924.2	731.3	369.3	644.7
7	19.30	116.8	73.74	48.18	21.00	20.71	520.4	977.4	922.4	730.5	369.3	641.8

3.2. Analysis of experimental results

As can be seen from Figure 1, the sales volume of Mosaic vegetables is the highest in the three years, while the sales volume of nightshade vegetables is the lowest. The sales distribution of each vegetable category in a day is consistent, that is, the sales peak of each vegetable is from 10 to 11 and 17 to 18, and the sales gradually decline after 19 o 'clock.

Through the analysis of the fitting graph of each category, it can be concluded that the sales volume of each category decreases with the increase of pricing. Therefore, in this study, Mosaic vegetables with better fitting degree were selected for time series prediction. Comprehensive analysis is made according to Figure 3, Figure 4, Table 1 and 2, based on the correlation variables, it can be concluded from the Q statistic results that Q6 does not show significance in level, and the hypothesis that the residual of the model is a white noise sequence cannot be rejected. Meanwhile, the goodness of fit R^2 of the model is 0.497 and 0.633 respectively, which indicates that the model performs well and basically meets the research requirements. Therefore, for all Mosaic vegetables, the total replenishment volume is about 99.3kg, and the total price is about 992.7 yuan /kg. On this basis, considering the reasonable float of different specific individual varieties, the pricing strategy can be completed.

4. Conclusions

For most application fields, it is an important means to reveal the internal relationship of variables to establish proper mathematical function model fitting the relationship between variables based on existing data. After pre-processing a large amount of data collected, this paper visualized the data, established the distribution map of vegetable sales in recent years, analyzed the best-selling vegetable

varieties, and the chart was clear and intuitive. Then, the linear fitting of the data reveals the relationship between pricing and vegetable sales. The sales volume of each category decreases with the increase of pricing. The model is simple and easy to understand. Through the ARIMA time series forecasting model, the quantity of vegetable replenishment and pricing in a certain period of time were obtained. Due to the instability of pricing and sales volume, the forecasting degree performed well within the error range. This study meets the future market demand and consumption level, is conducive to maintaining the interests of agricultural producers, has great guiding significance for the development of market economy, and has important reference value for the data analysis and formulation of optimization programs in the vegetable industry.

References

[1] Fuhua Shu. Prediction of intelligent lighting market size in China by Long- and Short-term memory neural Network [J]. China Lighting Electric Appliances,2023(06):16-20.

[2] Cameron Adrian J,Brown Amy,Orellana Liliana,Marshall Josephine,Charlton Emma,Ngan Winsfred W,Ananthapavan Jaithri,Isaacs Jasmine,Blake Miranda,Sacks Gary. Change in the Healthiness of Foods Sold in an Australian Supermarket Chain Following Implementation of a Shelf Tag Intervention Based on the Health Star Rating System. [J]. Nutrients, 2022, 14 (12).

[3] Danlin Wei. Research on risk and pricing of vegetable income insurance in Shandong Province [J]. Rural Economics and Science and Technology,2021,32(02):69-70+160. Central South University of Forestry and Technology,2023.

[4] Anpin Wei, Lin Zhang, Cuncun Wang, Zengfei Wang, Haosen Zhao, Xiaodong Wang. Construction and demonstration of agricultural product price index [J]. Journal of Library and Information Technology in Agriculture :1-9.

[5] A Yang, Pengkun Xu. Fault Location of Ground cable overhead hybrid line based on time linear fitting [J]. Journal of Chongqing Technology and Business University (Natural Science Edition):1-10.

[6] Lingling Yang, Xiangchun Ma, Minggao Su. Regression analysis of engineering cost of nuclear power project under "dual carbon" target [J]. Industry and Technology Forum, 2019,22(02):57-59.

[7] Yujie Wang, Li Bai, Banghong Zhao. Prediction of garlic price in China based on ARIMA model [J]. Vegetables,2021(12):50-54.

[8] Wenhua Zhang. Algorithm Improvement and simulation of ARIMA model for time series prediction [J]. Information and Computer (Theoretical Edition), 21,33(05):53-56.

[9] Detian Bu, Long Chen, Hongbin Liu, Yijie Cheng, Yinian Han. Weibull parameter Test based on vibration time Series [J]. Forestry Machinery and Woodworking Equipment :1-12.