# Research on Influencing Factors of Fruit Futures Turnover Rate and Prediction Based on Random Forest Regression Model

Rutang Yuan[1, a]

[1]College of Science, Wuhan University of Technology, Wuhan, Hubei, 430000, China

[a]Corresponding Author's Email: 2647477276@qq.com

## Abstract

**Under the influence of the epidemic, the changes in the turnover rate in 2020 was different. This paper uses auto-correlation diagrams and white noise tests to analyze the change in the turnover rate of domestic single fruit futures in the past one year from the beginning of 2020 to the beginning of 2021 and the impact of the epidemic. The study found that the change in turnover rate is not cyclical, but has a linear trend, the data is internally correlated, and does not have random fluctuations, and the turnover rate was greatly affected by the epidemic in 2020. When the epidemic was the most serious, fruit futures' price trend was poor. When the impact of the epidemic became smaller, the price trend of fruit futures rebounded greatly. And use the random forest algorithm and Kruskal-Wallis H test method to get the significant and non-significant influencing factors of the fruit futures turnover rate. According to the obtained significant factors,we use random forest regression to predict the turnover rate random forest. The regression prediction model can be obtained through inspection. The prediction result of the model is better and the error is small . Reasonable use of the prediction model can bring huge benefits to enterprises and individuals. At the same time, by comparing the multiple linear regression forecasting model, the ARIMA model and the random forest forecasting model established in this paper, it is found that the random forest regression forecasting model established in this paper has higher accuracy and is more suitable for analyzing the influencing factors of turnover rate and its impact on its predictions.**

## Keywords

**Turnover rate; Random forest regression; Kruskal-Wallis H test; ARIMA model.**

## 1. Research Background and Significance

A sudden epidemic in early 2020 had a huge impact on the futures market, and the turnover rate was a barometer of the popularity of futures performance. Turnover rate[1] refers to the frequency of futures changing hands in the market within a certain period of time, and is one of the indicators reflecting the liquidity of futures. The stock index futures market is an unstable, open, non-linear dynamic changing complex system. The changes of futures contract prices in the market are affected by many factors such as finance, economy, politics, society and investor psychology. Generally speaking, the higher the turnover rate, the more the product is favored by consumers. Therefore, it is particularly important to find the influencing factors of the turnover rate and predict its future development. If the prediction model is better, it can be used for enterprises and individuals to bring huge benefits.

In the research on the determination and prediction of the influencing factors of China's futures market, many literature are studying the changes of futures prices and do not analyze and predict the changes in the turnover rate, but directly predict from the changes in prices ( For example, Li[2] uses the change data of futures prices to use BP neural network to predict futures prices; Chen[3] uses machine learning to predict bitcoin futures prices), and analyzes its influencing factors from futures prices (such as Tao[4] change studies have drawn relevant conclusions that are largely influenced by speculative factors and investor sentiment). However, these literature have not analyzed the influencing factors and predictions of the turnover rate, an important indicator, especially in the fruit futures market. There is no literature to study the influencing factors and changes of the turnover rate in China's fruit futures market, and make reasonable predictions. Fruit is a daily necessity of our life, and the price changes in the futures market are closely related to our lives. The changes in the turnover rate have an important impact on the futures price trend. Therefore, the research forecast on the changes in the turnover rate of China's fruit futures market is significant. It is necessary that a comprehensive analysis of influencing factors and a better prediction model can bring significant benefits to fruit manufacturers and individuals.

## 2. Materials and Methods

### 2.1. Research Object and Source

This study will study the relevant data of a certain fruit futures to obtain its influencing factors and give the corresponding prediction model.

The futures-related data of this research comes from the daily relevant data of a domestic fruit futures.

The futures trading data is mainly from a commodity exchange, including the daily turnover rate, change rate and range of the fruit futures from February 3, 2020 to January 5, 2021 , margin, service fee, closing positions on the day, whether there are severe weather conditions, whether there are six factors indicators of peak consumption season.

### 2.2. Random Forest Algorithm

Random forest algorithm[5] is a statistical learning theory, which uses the bootsrap resampling method to extract multiple samples from the original sample, model a decision tree for each bootsrap sample, and then combine the predictions of multiple decision trees, and obtain the final result by voting. forecast result.

### 2.3. ARIMA Model

The basic idea[6] of the ARIMA model is that:With the change of time, the time-varying data formed by the predictor variables is regarded as a random sequence, and the sequence is described by a mathematical model. After the description is successful, the model can be used for prediction, and the past value of the described object can be used to predict the future value.

Based on the data given, this paper studies the change law of the turnover rate from four aspects: data cycle, data trend, data randomness, and data internal transitivity. Then, the influencing factors are analyzed, the random forest algorithm is used to rank the feature importance of each factor, and the Kruskal-Wallis H test method is used to comprehensively judge whether these factors have an impact on the turnover rate. Finally, the obtained results are summarized and analyzed, and the factors that have a significant and no significant effect on the turnover rate are obtained. Then use the random forest regression model to make predictions. This paper also discusses the multiple linear regression forecasting model and the time series forecasting model based on turnover rate, and makes a comparative analysis with the established random forest forecasting model.

# 3. Results

## 3.1. Analysis of the Changes in Turnover Rate

Continue to explore the law of changes in the data. First, the time series diagram of the original data of turnover rate and its auto-correlation diagram are shown in Figure 1 and Figure 2.

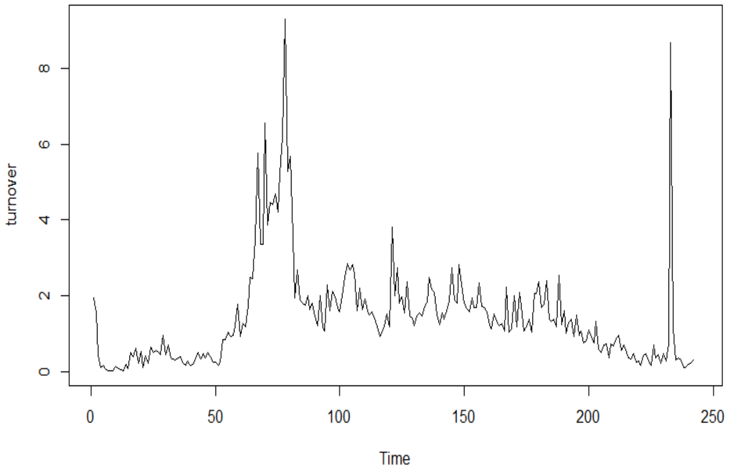

**Figure 1.** Timing diagram of original data

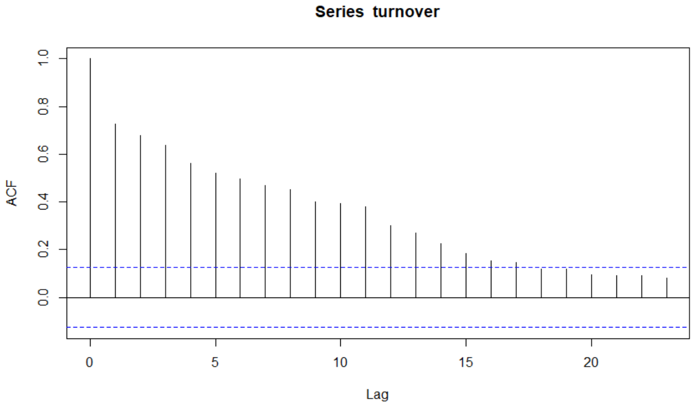

**Figure 2.** Image of auto-correlation coefficient of original data

Observing Figure 1,it can be seen that the turnover rate data was relatively stable from February to April 2020, and fluctuated between 0% and 1%; it showed a sharp upward trend in May 2020, and reached a peak of 9.29 on May 20, 2020. %, and then declined; it was relatively stable from June to October 2020, fluctuating between 1% and 3%; it gradually decreased from October 2020 to the beginning of January 2021, but it briefly appeared on December 23, 2020. The maximum value is 8.68%.

The new crown epidemic was widely spread in January 2020. Through the changes in turnover rate, it can be found that in February, when the epidemic was the most rampant, the turnover rate indicator dropped to the lowest in the whole year. It can be judged that the epidemic has a greater impact on the futures economy, but With the continuous efforts of the government and the people, it can be found that in April, the epidemic gradually dissipated and the economy gradually recovered. It reached its maximum value in mid-May and the economy developed rapidly. However, due to the impact of the epidemic, the futures economy changed. The rate has dropped.

At the same time, it can be roughly judged that the change trend of the turnover rate is not periodic. It can be seen from Figure 2 that with the increase of the time shift length, the auto-correlation coefficient has been decreasing, and there is no periodic change, so the change of the turnover rate The change is not periodic, and most of the auto-correlation coefficients are larger than the endpoints of the confidence interval, and the degree of change is large, so the change of the turnover rate is not stable.

The original data is not stationary. For non-stationary time series, the stationary data can be obtained by difference, and the first-order difference is performed on the original data to obtain the time series diagram and auto-correlation diagram after the first-order difference of the original data, as shown in Figure 3 and Figure 4.

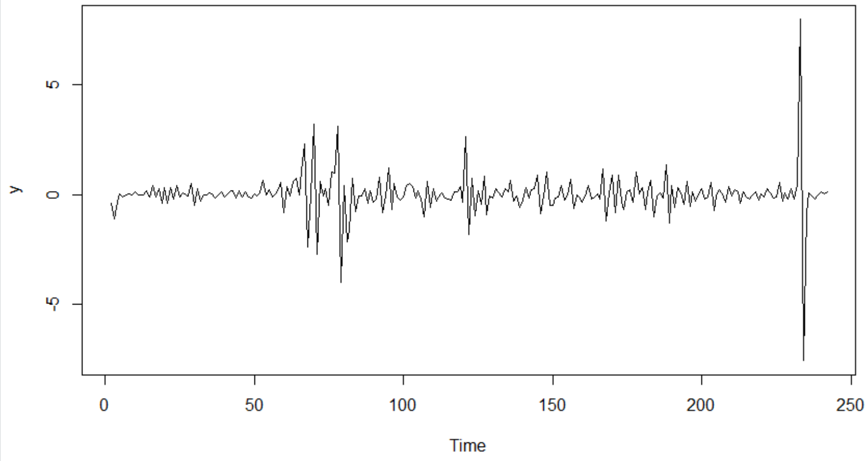

**Figure 3.** Timing diagram of first-order difference data

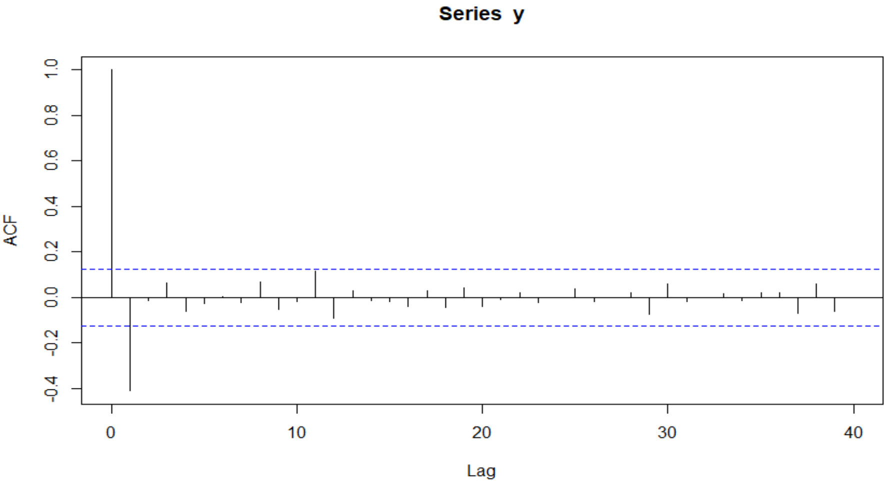

**Figure 4.** Auto-correlation coefficient diagram of first-order difference data

It can be seen from Figure 3 and Figure 4 the timing diagram is relatively stable. Further observation of the auto-correlation diagram shows that when the time translation length is small, the value of the auto-correlation coefficient is larger, and when the time translation length expands,the value of the auto-correlation coefficient is generally approaching 0. The correlation coefficient is tailing, so the data is stable.

This data is obtained from the first-order difference of the original data, indicating that the original data has a linear trend.

The data after the first-order difference is stationary, and the randomness of the data and the internal transitivity of the data can be judged by the white noise test. The p-values of the 6th-order delay, the 12th-order delay, and the 18th-order delay are all less than 0.0001 and far less than 0.05, so there is a correlation between the data. The data is non-white noise after the difference operation, so the data does not have randomness fluctuation.

### 3.2. Conclusion of the Changes In Turnover Rate

Based on the above analysis, the following conclusions are drawn: the change of turnover rate is not cyclical, but has a linear trend, the data is internally correlated, and does not have random fluctuations, and the turnover rate was greatly affected by the epidemic in 2020.

## 4. Analysis of Influencing Factors

In the following, the random forest algorithm is used to obtain the feature importance ranking of the six factors, and then the Kruskal-Wallis H test[7] method is used to judge the influence of the features on the turnover rate. Finally, the two methods are combined to analyze the factors that affect the turnover rate and those that do not.

### 4.1. Random Forest Algorithm for Feature Importance Ranking

Feature importance evaluation[8] using random forest.

The variable importance score is used to represent, and the Gini index is represented by GI. There are 6 features$X_1 \dots X_6$, representing factors of ups and downs, margin factor, handling fee factor, day closing factor, whether there is a disaster weather factor, whether there is a peak consumption season factor.Now, the Gini index score $VIM_j^{(Gini)}$of each feature $X_j$ is calculated and normalized, that is, the average change of the j-th feature in the node splitting impurity of all decision tree species in the random forest.

From this, the feature importance ranking is obtained, as shown in Table 2 below:

**Table 1.** Ranking table of the importance of influencing factors

| Influencing factors | feature importance |
|---|---|
| ups and downs | 0.620798 |
| Margin | 0.230818 |
| handling fee | 0.119973 |
| Close position on the day | 0.011414 |
| Is there any disaster weather | 0.009928 |
| Is it a peak season for consumption | 0.007069 |

It can be seen from Table 1 that the importance of the ups and downs factors is greater than 0.25, which is 0.620798. Therefore, it can be obtained that the ups and downs factor is a more significant factor at this time.

### 4.2. Kruskal-Wallis H Test to Screen for Significant Factors

Because the Kruskal-Wallis H test can only be used when the data is an ordered categorical variable, the ups and downs factors and margin factors cannot be judged by this method, and the H test can be performed on the remaining 4 features, as shown in Table 2 below.

**Table 2.** H test result table

| | Margin | handling fee | Close position on the day | Is there any disaster weather | Is it a peak season for consumption |
|---|---|---|---|---|---|
| test statistics | 65.688 | 7.944 | 57.163 | 2.162 | 6.657 |
| degrees of freedom | 4 | 2 | 3 | 1 | 1 |
| P value | <0.0001 | 0.019 | <0.0001 | 0.141 | 0.010 |

The P values of the four factors, namely the margin factor, the handling fee factor, the day closing factor, and whether the consumption peak season factor is less than 0.05. The sample distribution is different, so it is considered that the above four factors have no significant impact on the turnover rate; and the P value of whether there is a disaster weather factor is 0.1414 greater than 0.05, accept the null hypothesis, and think that the sample distribution is the same as that of the turnover rate. Therefore, It is believed that whether there is a disaster weather factor has an important impact on the turnover rate.

To sum up, it can be concluded that the fluctuation factors and whether there are disaster weather factors have an vital impact on the turnover rate of fruit futures.

## 5. Random Forest Regression Prediction Model

The following is a prediction model of turnover rate based on the random forest algorithm, and the random forest algorithm is used to solve the regression problem.

According to the conclusions obtained in Section 3, the ups and downs factors and whether there are disaster weather factors have a significant impact on the turnover rate. Therefore, for this problem, we use the data of these two characteristics to establish a random forest model to perform regression prediction on the turnover rate. Based on the random forest model established in 3, the random forest model is established by using the data of the fluctuation factor and whether there is a disaster weather factor. Finally, the data of the two factors can be input, and the corresponding turnover rate can be directly obtained.

After establishing the random forest model for these two factors, enter the data of the last two days of influencing factors, and get the following table 3.

**Table 3.** Prediction table of turnover rate

| Date | Turnover forecast |
|---|---|
| 2021/1/6 | 0.99956255 |
| 2021/1/7 | 1.11858981 |

For the test of the established random forest model, it is found in 3 that the fluctuation factors and whether there are disaster weather factors have a significant impact on the turnover rate. The establishment of the random forest model for these two data is tested, and the final set is $R^2$ =0.83627>0.8, which proves that the model works well.

## 6. Discussion

Considering the influence of all factors, a multiple linear regression model is established to compare with the above model to judge the accuracy of the model. At the same time, without considering the influence of factors, the ARIMA time series model is established based on the turnover rate alone, and the accuracy analysis is carried out with the original prediction model.

## 6.1. Multiple Linear Regression Prediction Model

Use Rstudio software to perform multiple linear regression analysis, set the turnover rate as y, and set the change rate, margin, service fee, position closing on the day, whether it is a severe weather condition, and whether it is a peak consumption season to$x_1$, $x_2$, $x_3$, $x_4$, $x_5$, $x_6$.

Get its model as follows:

$y = 2.354 + 0.000x_1 \text{-} 12.780x_2 + 0.024x_3 + 0.020x_4 + 0.529x_5 \text{-} 0.650x_6$

Its R square is 0.123, which is far less than 0.8, and the fitting effect is very poor, so the multiple linear regression prediction model is not suitable.

Comparing multiple linear regression and random forest models: The goodness of fit of random forest is generally higher than that of multiple linear regression, which improves the prediction accuracy without significantly increasing the amount of computation. Random forest is better at dealing with multi-collinear data, and it is not easy for multi-collinearity to lead to unstable regression coefficients, which seriously affects the accuracy of the model. Random forests are better at dealing with interactive data. So using random forest algorithm is far superior to multiple linear regression.

## 6.2. ARIMA Model

For the selection of the prediction model, in addition to the random forest model, we can also use the ARIMA time series model to predict the turnover rate. Use Rstudio to build the time series model of turnover rate. For the turnover rate, combined with the AIC information criterion (the lower the value, the better), after comparison and selection, the optimal model is finally found: ARMA(0,1,1 ), whose model formula is:

$y(t) = \text{-}0.005 \text{-} 0.540 * \varepsilon(t\text{-}1)$

The model parameter table is Table 4 below:

**Table 4.** Model parameter table

| ARIMA(0,1,1) model parameter table | | | | | | |
|---|---|---|---|---|---|---|
| item | symbol | coefficient | standard error | z-value | p-value | 95% CI |
| Constant term | c | -0.005 | 0.026 | -0.176 | 0.860 | -0.056~0.046 |
| MA parameters | β1 | -0.540 | 0.059 | -9.175 | 0.000 | -0.656~-0.425 |
| AIC value: 626.204<br>BIC value: 636.659 | | | | | | |

The fitting forecast of the turnover rate time series model is shown in Figure 5:

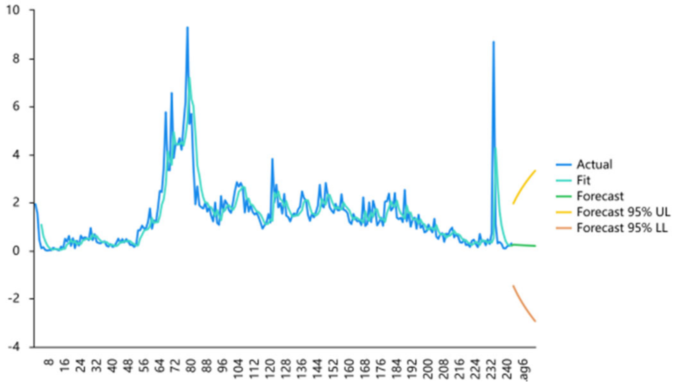

**Figure 5.** Model fitting prediction diagram

The forecast values for the next six periods are shown in Table 5 below:

**Table 5.** The prediction results of the ARIMA model for 6 periods backward

| predict | Predicted turnover rate |
|---|---|
| 1 period backward | 0.248 |
| 2 period backward | 0.243 |
| 3 period backward | 0.239 |
| 4 period backward | 0.234 |
| 5 period backward | 0.229 |
| 6 period backward | 0.225 |

The forecasts made by the ARIMA time series model have small fluctuations. There are still some problems: if the data itself has a decreasing trend, and the fluctuation range after the difference is relatively small, then the long-term forecast will definitely still have a negative value. This problem shows that it is impossible to impose constraint resolution, because ARIMA itself does not have this processing mechanism.

## 7. Conclusion

Firstly, this paper studies the changing law of turnover rate from four aspects: period, trend, randomness and internal transitivity of data. The final conclusion is that the turnover rate is not cyclical, has a linear trend with time, does not have random fluctuations, and has internal correlations in the data. The turnover rate will be greatly affected by the epidemic in 2020. The following uses the random forest model to evaluate the importance of influencing factors (features) and the Kruskal-Wallis H test method to obtain the factors that have a significant impact on the turnover rate. Later, the prediction of the turnover rate is carried out. According to the factors with significant influence obtained above, the random forest regression prediction modeling is carried out with the fluctuation factor and whether there is a disaster weather factor as the independent variable, and the turnover rate as the dependent variable. The model test shows that the random forest model established in this paper has a high goodness of fit ($R^2$ =0.83627) and a small error. Finally, the accuracy of the multiple linear regression forecasting model and the ARIMA time series forecasting model is discussed, and a comparative analysis is made with the random forest forecasting model. It is comprehensively judged that the random forest regression forecasting model established in this paper is more practical.

## References

[1] Zeng Yan.(2015). The Dictionary of Management. Dong Yunhu, editor-in-chief (eds.) Shanghai Yearbook. Editorial Department of Shanghai Yearbook, 311-312.

[2] Li Cong.(2012). Prediction of stock index futures price based on BP neural network (Master's thesis, Qingdao University).

[3] Chen Zheshi.(2021). Research on the Internet and Financial Futures Influencing Factors and Prediction Methods of Bitcoin Price (PhD Thesis, Harbin Institute of Technology)

[4] Tao Libin, Pan Wanbin & Huang Junzhe.(2014). Changes and determinants of price discovery ability of CSI 300 stock index futures. Financial Research (04), 128-142..

[5] Ma Jinsha. (2021). Variable screening method based on random forest variable importance score and its application in tumor typing and diagnosis (Master's thesis, Shanxi Medical University).

[6] Chen Kexiu & Liu Juan.(2022). Sales Forecast of Euler Black Cat New Energy Vehicles Based on ARIMA Model. Modern Industrial Economy and Informatization (03), 169-171.

[7] Qi Haiping & Shen Xiping.(2015). Implementation of Kruskal-Wallis H test for multiple comparison of mean rank in SPSS software. Journal of Lanzhou Institute of Technology (02), 76-78.

[8] Wu Chao & Luo Jing.(2018). Automatic identification of tax evasion based on random forest. Software Guide (08), 13-16.