Forecasting Liquor Index: Applications of ARIMA and Gray Model

Haoran Tan[1, *, †], Jiayang Gao[2, *, †], Duping Gao[3, *, †]
[*]danieltannnnnn@gmail.com, [*]jgao68@emory.edu, [*]dtgao@emory.edu

[1]School of Economics and Management South China Normal University Guangzhou, Guangdong, China

[2]Emory College of Arts and Sciences Emory University Atlanta, Georgia, the United States

[3]Emory College of Arts and Sciences Emory University Atlanta, Georgia, the United States

[†]These authors contributed equally.

Abstract—An imperfect stock market provides ambitious investors with plenty of room for arbitrage. Currently, ARIMA (Autoregressive Integrated Moving Average) and Gray forecast models are widely used to forecast future stock prices. In our paper, we aim to investigate the efficiency of those two models. Since each stock is idiosyncratic by nature, it is not advisable to look for one single forecast method that is robust for forecasting all the stocks in the market. Thus, we select the Chinese liquor industry, one of the most popular industry to invest in for the past year in China, as the context to study these two models. To compare the efficiency of those two models for forecasting the Chinese liquor index specifically, we first format the data into the appropriate form and then test various assumptions hidden in ARIMA and Gray forecasting models. After applying ARIMA and Gray forecast models respectively to predict stock closing prices for 20 days, we conclude that the ARIMA model performs better with a smaller sum of squared residuals. As a result, we deduce that ARIMA has a better ability of forecasting Liquor Index than Gray model. Methods incorporating more factors to forecast index deserve further investigation.

Keywords: ARIMA, Gray Forecasting, Liquor Index

1 INTRODUCTION

Stock price forecasting is essential for individual and institutional investors. The price fluctuations of stocks and related indices bring profits to investors directly in the capital market. According to the efficient market hypothesis, our capital market is roughly between a weak efficient market and a semi-efficient market, which allows for opportunities for excessive returns. The impact of industry on stock prices is significant. The existence of problems and future trends of an industry determines the stock price trend of that industry. Lv Junzuo argues that different industries have different returns in various stages of the economic cycle [1]. The Shanghai and Shenzhen markets are divided into sectors including traditional agriculture, manufacturing, transportation, education, food and beverage, finance, and culture and sports. There are sub-sectors under the major sectors. Among them, since the middle of 2020, the chemical, electrical equipment, automotive industry, non-ferrous metals, and food and beverage industries have outperformed, with electrical equipment rising even close to 100%, non-ferrous metals reaching about 60%, and food and beverage reaching 34.37%. Dong Lingyu (2019) studied the impact of

ICEMME 2022, November 18-20, Nanjing, People's Republic of China

DOI 10.4108/eai.18-11-2022.2326834

supply-side reform on stock prices in the chemical industry [2]. The infrastructure construction and energy economy transformation on the technology side led to the rapid development of the electrical industry with better prospects for the capital market [3]. Tang Baojun et al. argue that the automotive industry is developing toward new energy and "vehicle-electricity separation," with excellent market potential [4]. Dan Lai analyzed the non-ferrous metals industry from the industrial chain perspective and finds that the problem of a low surplus of enterprises can generate more profit if optimized [5]. Wang Yiwei investigated the food and beverage industry for its relatively stable growth, in which the liquor branch performs well and ranks first in all core indicators [6]. Compared to other sub-divisions of food and beverage such as spices and soft drinks in the red sea track, the liquor industry can be described as unique. Chen Jingru analyzed the liquor industry's market structure, behavior, and performance and proposed the management of consumer groups and marketing channels [7]. Wang Xin et al. proposed modifications in the profitability model of liquor to improve the profitability status of the whole liquor industry [8]. Qin Weiyao used the EVA model to evaluate the value of listed companies in the liquor industry for research [9], and Zhang Renping employed factor analysis to evaluate the performance of the liquor industry [10]. However, there is a gap in the existing literature concerning the content of stock price forecasts and related guidance recommendations for investors in the liquor industry.

The methods for stock price forecasting generally include fundamental, technical analysis, statistical methods, and neural network forecasting methods, with the latter two being studied in greater greater details in the last two years. Li Zhengrong used a combination of weighted support vector machine and relief algorithm to predict the trend of stock data in six industries with an accuracy of more than 70% [11]; Gupta improved Markov Model and compared the improved model with ARIMA, and ANN (artificial neuron network), which paved a new way for stock price forecasting [12]. In terms of neural network prediction, Zhang Ni used an LSTM (Long Short-term Memory) model to predict the stock price of Guizhou Maotai with a good fit [13]. Tang Jianqing used a trained BP neural network to select a more accurate model parameter function for prediction and eventually used it for reference in the investment strategy of individual stocks in the industry [14]. Most promising ARIMA models for predicting stock prices are only simple forecasts for a particular market index. Monal and Adebiyi found that ARIMA has better stock price forecasting ability in the US, Nigeria Indian stock market, respectively [15,16]. While grayscale forecasting models have been widely studied in the field of stock price forecasting, Chang used the corrected grayscale model and Garch to forecast the NYSE and NASDAQ markets [17], and Xiao used the error corrected grayscale forecasting model WGM for forecasting with promising results [18]. Consequently, existing literature only studies ARIMA and gray forecasting models in separated settings, resulting in the absence of objective cross-sectional comparison of the two models. This study aims to address this issue by providing a close comparison of these two models.

In this paper, we choose to focus on the more common ARIMA model and Gray forecasting model in statistical methods to study the Chinese liquor index, which reflects the boom level of the liquor industry, and then compare the accuracy of the two models for stock forecasting. The historical data of the liquor index is put through the ARIMA model and the Gray forecasting model to derive the results respectively and compared with the actual data to evaluate the effectiveness of those two models.

2 METHOD

In this paper, we forecast the data set as the Chinese liquor index (SCN: 833137.EI) from December 30, 2019 to March 5, 2021 mainly by processing the trading date and closing price using ARIMA and Gray forecasting models respectively for the next 20 days of data and then comparing the forecast values with the actual data.

The ARIMA model, which integrates auto-regressive and sliding average, is a method for predictive analysis of time series data. The model is widely used in fields such as business, medical, transportation, and social science and plays a vital role in forecasting market demand, commodity prices, and future trends in the industry. Steps used in the model are data acquisition, preprocessing, model testing, order fixing, estimation of parameters, and model validation.

Gray forecasting model has a different logic, which derives from the notion of the White System and the Black System. The White System means that the internal characteristics of the system are entirely known; the Black System means that the internal information of the system is wholly unknown, and the Gray System is a system between the White System and the Black System. Part of the internal information of the Gray System is constant, while the other part of the information is unknown or uncertain. Gray Forecasting refers to predicting the development and change of the characteristic value of system behavior, the prediction of the system that contains both known and uncertain information. That is, the gray process that changes within a specific range and is related to a time series. Although the phenomenon shown in the gray process is random and chaotic, it is still orderly and bounded, after all, so the dataset obtained follows lurking laws. Gray Prediction is to use this law to establish a Gray Model to quantify the uncertainties and employ the known information to find the pattern of motion of the designated system. At present, the most widely used Gray Forecasting model is the GM(1,1) model of a variable and first-order differential for sequence forecasting. It is based on a random original time series. The new time series formed by time accumulation can be approximated by the solution of a first-order linear differential equation. And it has been proven that when the time series implies the exponential change law, the prediction of the Gray Model GM(1,1) is very successful.

2.1 ARIMA model

2.1.1 ARIMA model building

The ARIMA model is a combination of the autoregressive model AR and the moving average model MA. The autoregressive model uses data from the variable's history to make predictions about itself. In contrast, the moving average model is a re-optimization of the error term in the autoregressive model, with the equation mentioned below.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \in_t + \sum_{i=1}^{q} \theta_i \in_{t-i} \quad (1)$$

There are generally five steps to build an ARIMA model, five steps as mentioned above, and the process of each step is briefly described below.

Step 1: Obtain data from WIND and Eastern Wealth Choice, taking the key data column of the closing price. Although the trading dates are not continuous (every five days is a trading cycle), the closing price data can be consecutive. Since the dataset is large and the liquor index started a more apparent upward trend in January 2020, 284 data from the beginning of 2020 were extracted,

and the training and validation sets were divided in a 3:1 manner, with 213 data in the training set and 71 in the validation set. The missing values and outliers in the data were removed first.

Step 2: Model smoothness test. The Dickey-Fowler test is performed on the extracted data part, and the original hypothesis is that the time-series data are non-stationary. The value shows -0.76, which is greater than the critical value at the 1%, 5%, and 10% confidence level tests, indicating that the original hypothesis is accepted and the data are non-stationary. In addition, by customizing the function *test stationarity* and drawing the data trend graph, it is found that there are no more obvious temporal characteristics.

Table 1: Dickey-Fuller Test Result 1

Results of Dickey-Fuller Test:	
Test Statistic	-0.760516
p-value	0.830446
Number of Observations Used	284.000000
Critical value (1%)	-3.453587
Critical value (5%)	-2.871771
Critical value (10%)	-2.572222

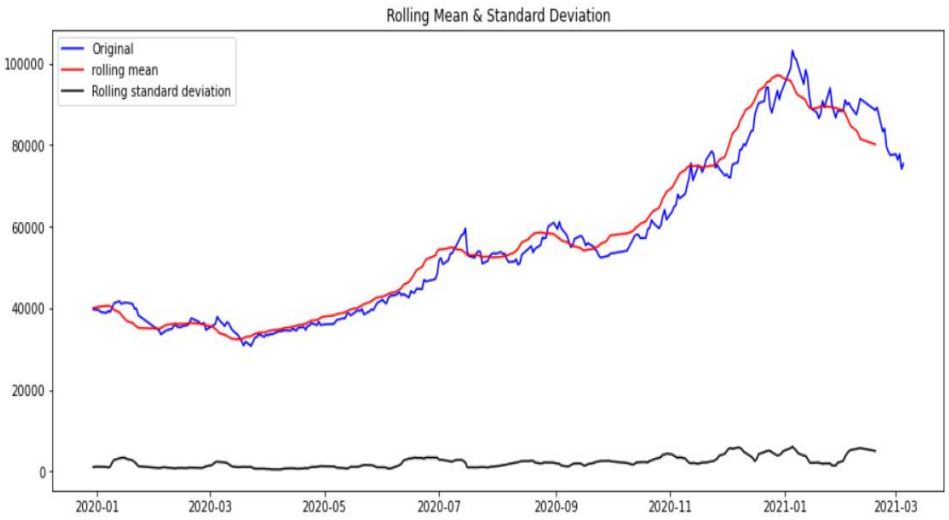

Figure 1: Rolling Means & Standard Deviation

To make the data reduce more seasonal factors influence, we use the first-order difference method; thus, the latest data are obtained as ts_log_diff. The DF test shows that the original hypothesis is rejected, indicating that the data are now smoother.

Table 2: Dickey-Fuller Test Result 2

Test Statistic	-1.449023e+01
p-value	6.165533e-27

Number of Observations Used	2.110000e+02
Critical value (1%)	-3.461727e+00
Critical value (5%)	-2.875337e+00
Critical value (10%)	-2.574124e+00

Step 3: Model Evaluation and Order Fixation. The parameters of (p, d, q) in the ARIMA model are estimated by ACF (Autocorrelation Coefficient) and PACF (Partial Autocorrelation Coefficient). Here, it is assumed that the number of lags is 20, the confidence level is 5%, and the p and q values are the values of the horizontal axis of the ACF and PACF images through the confidence interval, respectively, as shown below, and it can be seen from the figure as p and q may be equal to 1 or 2, respectively.

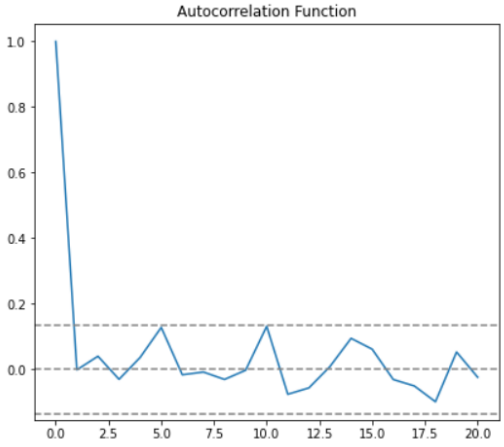

Figure 2: Autocorrelation Function

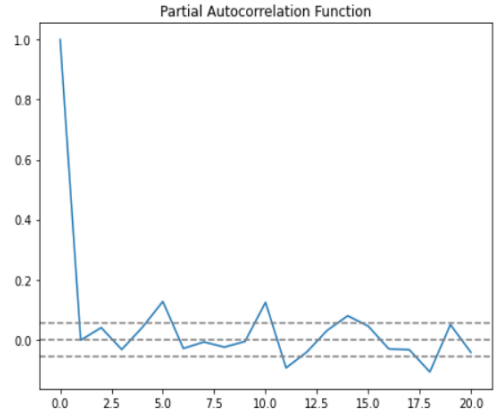

Figure 3: Partial Autocorrelation Function

Step 4: Model selection and testing. For the determination of p and q parameters, the adopted comprehensive consideration of AIC (Akaike Information Criterion), BIC (Bayesian Information Criterions), and HQIC (Hannan-Quinn information criterion) indicators, the four possible models are compared. Finally, the model with p and q of 1 is selected, at which time the three indicators are the lowest value respectively, and the optimum is reached.

Table 3: SARIMAX Results 1

Dep. Variable:	Closing price	No. Observations:	71
Model:	ARIMA(1, 1, 1)	Log Likelihood	-643.523
Date:	Sat, 31 Jul 2021	AIC	1293.046
Time:	10:57:51	BIC	1299.792
Sample:	71	HQIC	1295.725
Covariance Type:	opg		

Table 4: SARIMAX Results 2

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9364	0.061	-15.368	0	-1.056	-0.817
ma.L1	0.9087	0.08	11.3	0	0.751	1.066
sigma2	5.66E+06	9.37E-10	6.04E+15	0	5.66E+06	5.66E+06

Table 5: SARIMAX Results 3

Ljung-Box (Q):	19.92	JarqueBera(JB):	0.41
Prob(Q):	1	Prob(JB):	0.81
Heteroskedasticity (H):	1.07	Skew:	-0.06
Prob(H)(two-sided):	0.87	Kurtosis:	2.65

Comparing the trained ARIMA model with the original dataset, we can see that the fit is good and fits the actual data trend and values more closely.

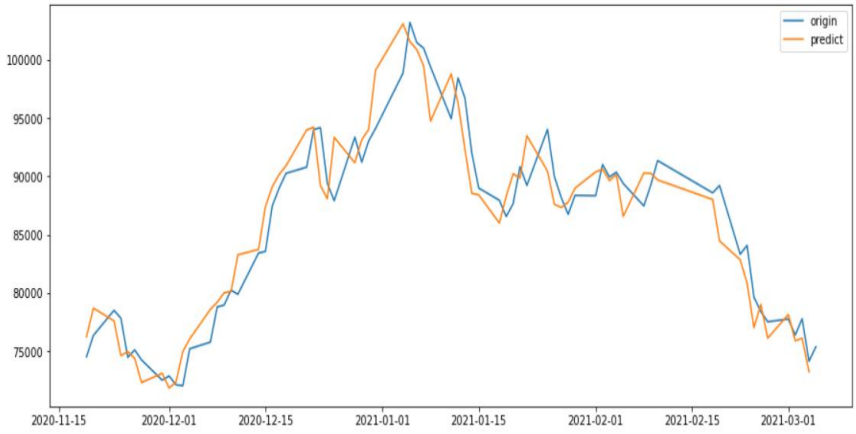

Figure 4: Comparison of origin and predict data

2.1.2 Model prediction and analysis

Based on the ARIMA model with p=d=q=1 and the original selection of 100 historical data, forecasts are made for the next 20 trading days to obtain the dataset NEXT. The results are kept for subsequent model comparison.

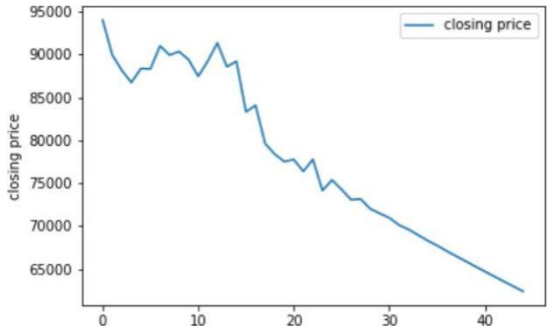

Figure 5: Forecast data by ARIMA

2.2 The Gray System

2.2.1 Gray model building

There are mainly two steps required to perform for the Gray model. The first step is to ensure the validity of our model. The second step can be used to predict the future closing prices by a sequence of mathematical manipulation.

Level Check: After storing data in a appropriate form, we check the level to ensure the validity of our model. We calculate its level by using the original data X(0) = (x(0)(1), x(0)(2), x(0)(3), · · · , x(0)(n)) and define $(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}$ where k = 2, 3, · · · , n. As long as for every k, (k) falls within (e−2/(n+1), e2/(n+2)), we pass the level check otherwise we need to do translation

transformation using Y (0) = X(0) + c to pass the level check. Fortunately, for our dataset, we pass the level check without any further transformation required.

GM(1,1) Model: First, we add the original cumulatively by defining X(1) = (x(1)(1), x(1)(2), x(1)(3), · · · , x(1)(n)) where $x^{(1)}(k) = \sum_{i=1}^{k} x^{(0)}(i)$ where k = 1, 2 · · · n. Then we define Z(1) = (z(1)(1), z(1)(2), z(1)(3), · · · , z(1)(n)) where z(1)(k) = 0.5x(1)(k) + 0.5x(1)(k − 1). By considering the Gray differential equation model x(0)(k) + az(1)(k) = b and subsequent equations like Shadow equation and Time response function, we have $\hat{x}^{(1)}(k+1) = [x^{(0)}(1) - \frac{b}{a}]e^{-ak} + \frac{b}{a}, \quad k = 1, ...n-1$. By subtracting $\hat{x}^{(1)}(k)$, we have our forecast function $\hat{x}^{(0)}(k+1) = \hat{x}^{(1)}(k+1) - \hat{x}^{(1)}(k) = [x^{(0)}(1) - \frac{b}{a}](1 - e^{a})e^{-ak}$, where $k = 1, ..., n-1$. To be consistent with the previous model, we predict the closing prices for 20 days using data from previous 25 days. Our result and the corresponding graph are shown below.

2.2.2 Gray Model prediction and analysis

By utilizing the Gray model GM(1,1) that we built, we intend to forecast the development of the liquor industry index for the next 20 trading days with 25 observations as the training set. As shown by the graph plotted, the GM(1,1) model forecasts a smooth, continuously decreasing trend for the index

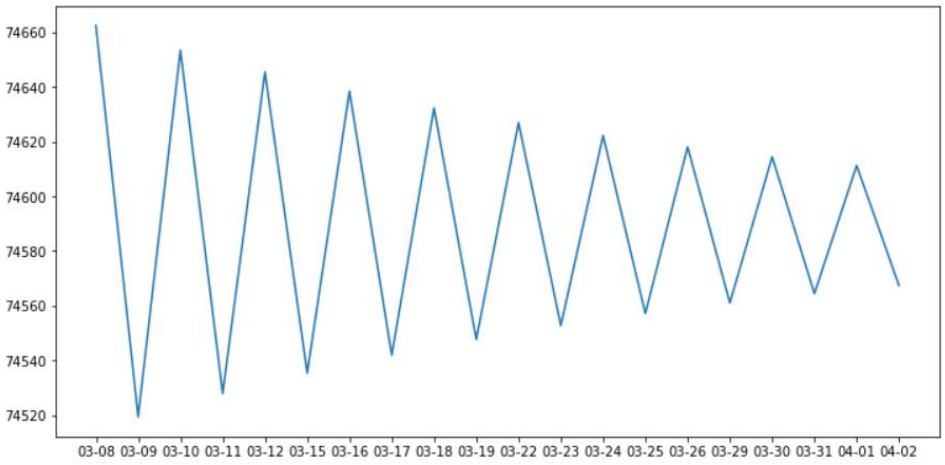

Figure 6: Forecast data by Gray Model

3 DISCUSSION AND COMPARISON

3.1 Discussion

From the ARIMA and Gray models, 20 forecast data were obtained for the forecast period from March 8, 2021 to April 2, 2021. The graphs comparing the forecast data with the actual data are shown below.

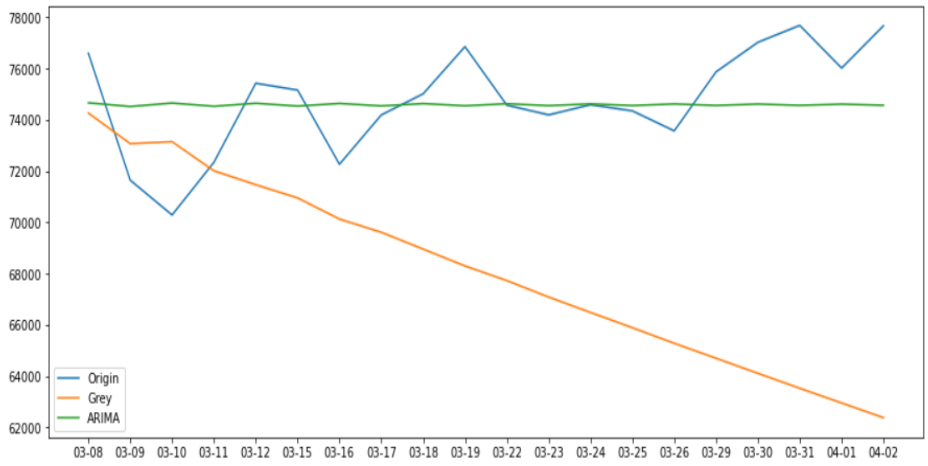

Figure7: Comparison origin data with Grey and ARIMA Model

The difference between the ARIMA model and the Gray prediction model and the actual data, i.e., the residual sum of squares, is calculated using the actual data origin as a benchmark, and the formula is shown below.

$$sum\_of\_residual = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad (2)$$

The sum of squares residuals of ARIMA is obtained as 1.38109, and the sum of squares of the Gray model is 7.82107. Thus, the residual sum of squares of the ARIMA model is much smaller than that of the Gray model, indicating that the model of ARIMA is better for forecasting the liquor index.

3.2 Comparisons

The ARIMA model uses a large amount of data, and the trend and seasonality of the time series data are well removed in the process of building, using, and testing the model. Although it is slightly tedious compared to the Gray forecasting model, the operation steps are relatively simple.

One characteristic of the Gray model is that it can use a smaller amount of data when forecasting. Since it is consuming to analyze a large amount of data to make the prediction, we take advantage of this characteristic and only use the closing prices for the previous 25 days to predict stock prices for the next 20 days.

For time-series stock data and data for the liquor industry index specifically, the ARIMA model has better forecasting results from the perspective of accuracy with a small sum of squares residuals.

4 CONCLUSION

The paper starts with stock price forecasting for the recently popular liquor industry. Using the liquor index 833137.EI as the data subject, the stock price forecasting is conducted using statistical methods rarely utilized in the liquor industry research field. Both models predicted the

data for the next 20 trading days, and a comparison with the actual residuals revealed that the sum of squared residuals was smaller for ARIMA. The liquor industry is cyclical in nature. The ARIMA model has high requirements for the smoothness of the data, and the medium- and long-term forecasts of the processed liquor index using the ARIMA model are more accurate.

The paper hopes to present the ARIMA model approach to providing some references and suggestions to investors in conducting investment activities. Market expectations for the liquor industry can be rationalized.

Due to the limited space, the paper fails to cover every detail of the model and data; it fails to consider more statistical models for comparison in the selection of statistical models; and for multiple liquor indices in the market, the paper only selects one of them as a representative, which fails to represent the whole liquor industry in a very scientific way. Last but not least, when evaluating the two models, we only consider the sum of squares residuals, which is not comprehensive. We plan to analyze more factors that are essential to models' precision and accuracy in the future.

REFERENCES

[1] Lv Junzuo. Research on stock industry selection and allocation based on economic cycle [D]. Nanjing University of Aeronautics and Astronautics, 2012.

[2] Dong Lingyu. Research on the impact of decapacity policy on stock prices in the context of supply-side reform [D]. Shandong Institute of Business and Economics,2019.

[3] " New infrastructure" to accelerate the conversion of old and new dynamics, the electrical industry to increase quality and speed [J]. Electrical Technology,2020,21(03):6.

[4] Tang Baojun, Wang Xiangyu, Wang Bin, Wu Yu, Zou Ying, Xu Huangchen, Ma Ye. Analysis and outlook of the development level of China's new energy vehicle industry[J]. Journal of Beijing University of Technology (Social Science Edition),2019,21(02):6-11.

[5] Lai D,Zhang Y. Research on the sustainability and influencing factors of surplus in nonferrous metal industry–based on the perspective of industrial chain[J]. Finance and Accounting Newsletter,2019(11):38-42.

[6] Wang Yiwei, Wang Hongyu. Outlook of brand development in food and beverage industry[J]. China Trademark,2019(06):14-18.

[7] Chen Jingru. SCP analysis of the liquor industry in Suqian City, Jiangsu Province [J]. Business and Management,2021,4(06):188-192.

[8] Wang Xin,Gulnar McBaity. Analysis of profit model of liquor industry[J]. Cooperative Economy and Technology,2021(01):108-109.

[9] Qin Weiyao. Research on the value evaluation of listed companies in the liquor industry: the example of enterprise W[J]. Old brand marketing,2021(06):109-110.

[10] Zhang Renping,Liu Junrong,Luo Jie. Evaluation of corporate strategic performance based on factor analysis method--the liquor industry as an example[J]. Business Economics,2016,35(02):80-84.

[11] Li Zhengrong,Wei Zengxin,Zhu Renjie. Research on stock forecasting based on Relief-WSVM[J]. China Management Information Technology,2020,23(11):150-152.

[12] Gupta, Aditya, and Bhuwan Dhingra. "Stock market prediction using hidden markov models." 2012 Students Conference on Engineering and Systems. IEEE, 2012.

[13] Zhang Ni. Research on the application of stock price prediction based on LSTM neural network[J]. Modern Business,2021(16):116-118.
[14] Tang Jianqing. Quantitative investment based on BP neural network[D]. Soochow University,2019.
[15] Mondal, Prapanna, Labani Shit, and Saptarsi Goswami. " Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices." International Journal of Computer Science, Engineering and Applications 4.2 (2014): 13.
[16] Ariyo, Adebiyi A., Adewumi O. Adewumi, and Charles K. Ayo. " Stock price Sprediction using the ARIMA model." 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation. IEEE, 2014.
[17] Chang, Ting-Cheng, Hui Wang, and Suyi Yu. " A GARCH model with modified gray prediction model for US stock return volatility." Journal of Computational Methods in Sciences and Engineering 19.1 (2019): 197-208.
[18] Xiao, Lifang, Xiangyang Chen, and Hao Wang. " Calculation and realization of new method gray residual error correction model." PloS one 16.7 (2021): e0254154.
[19] Information on https://blog.csdn.net/mengshangjy/article/details/79714747