Scientific Research Publishing

Open Journal of Business and Management, 2021, 9, 1312-1324
https://www.scirp.org/journal/ojbm
ISSN Online: 2329-3292
ISSN Print: 2329-3284

Precise Demand Forecast Analysis of New Retail Target Products Based on Combination Model

Jinli Jiang, Weiwei Yao, Xueyan Li

School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, China
Email: JinliJiang@163.com

How to cite this paper: Jiang, J. L., Yao, W. W., & Li, X. Y. (2021). Precise Demand Forecast Analysis of New Retail Target Products Based on Combination Model. *Open Journal of Business and Management, 9,* 1312-1324.
https://doi.org/10.4236/ojbm.2021.93071

Received: March 8, 2021
Accepted: May 28, 2021
Published: May 31, 2021

Abstract

In order to satisfy the consumers' pursuit of diversification of goods, new retail enterprises begin to gradually produce small quantities and various kinds of products, which makes the sales data become more complex and various, and then makes the inventory management more difficult. Therefore, it is very necessary to establish an accurate demand prediction model for the sub-category stratum. In this paper, we firstly consider the effect of external macrofactors on sales, and establish a multiple linear regression model to forecast the sales of the target products. Then we consider the regularity and trendency of previous sales, comparing the fitting degree of different parameter ARIMA models, and finally establish the ARIMA (2, 2, 1) model with the best prediction effect. Finally, in the light of the fitting degree, the two models are given different weights, and a predictive model that combines multiple linear regression and ARIMA (2, 2, 1) is established. It can be shown from the results that the prediction effect of combined model is better and it can accurately predict needs for new retail goods, thereby reducing the difficulty of inventory management and improving corporate competitiveness.

Keywords

Demand Forecast, Multiple Linear Regression, ARIMA Model, Combination Model

1. Introduction

In the context of the quick increase of the Chinese commodity economy and the comprehensive popularization of Internet technology, new retail enterprises which combine the Internet technology, big data technology and logistics tech-

nology emerge as the times require. However, physical retail industries, which take commodities as the core and only focus on the inventory management of commodities, cannot adapt to the digital era and fully satisfy the demands of consumers, new retail enterprises are different from the traditional retail enterprises, on the one hand, new retail enterprises are the combination of e-commerce platforms and store scene consumption. It brings together consumers from multiple sales channels such as online e-commerce and physical stores, through online platform construction and offline immersive scene consumption, to provide consumers with full service and increase consumers' shopping experience. On the other hand, it is more humanized and more focused on the service to consumers, and the core of the business is transformed from the previous commodity to the commodity plus service. With the increase of people's income and the great abundance of material, the consumption willingness and consumption level of residents are also improved, and the demands of consumers have various types. New retail enterprises use big data mining technology, combined with consumers' hobbies, behaviors, habits and other aspects of user characteristics, continuously to improve the production model, further subdivide the product hierarchy, and produce more diverse, beautiful and fashionable target products to satisfy the diverse, fashionable, and personalized demand of consumers. Although this production mode can serve consumers better, predicting consumer needs is difficult when the sales data is complex, which also leads to a variety of challenges, such as the production plan is difficult to formulate, inventory is hard to administer and so on. Therefore, considering the effect of external macro factors and the regularity and trend of historical sales data, this article builds a model on the basis of the multiple linear regression and ARIMA (2, 2, 1) in order to provide a more accurate demand analysis and sales forecast for regional level, sub category level and even store skc level, and further make inventory management simple and enhance the profitability and competitiveness of new retail enterprises.

2. Literature Review

(Gong & Huang, 2017) combined grey theory and exponential smoothing method to establish a model to predict product demand. However, the gray theory is not good for long-term prediction and is only sui for small samples. (Miao, Tang, & Luo, 2020) used the ARIMA model to forecast the sales of new energy vehicles, taking into account the seasonal factors of historical sales data. (Dong, Dong, Zhang, & Cui, 2020) used the redesigned traditional data as the actual input of the exponentially weighted average method, which improved the accuracy of corporate sales forecasts. (Rong & Guo, 2019) used convolutional neural networks to predict online product sales, taking into account external factors affecting online product sales. (Wu, Lin, Li, Wu, Wang, & Wu, 2016) used the support vector machine model to predict cigarette sales, and studied the non-linear relationship of cigarette sales data. (Liang, 2018) proposed a combined model based on the FBProphet model and the LightGBM model to predict

hotel online sales. The results show that the combined model has higher prediction accuracy. (Zhang & Qiu, 2019) used the decision tree model that can effectively deal with the problem of nonlinear regression to predict the sales of gas stations, and obtained a good prediction effect. (Wang, 2019) established a forecasting model through factor analysis of the sales data of heavy trucks. (Zhang, 2020) proposed a combined model based on ARIMA time series and BP neural network to study both the linear and non-linear characteristics of dish sales data. (Yang, 2017) established multiple linear regression and BP neural network models to predict the passenger car market by analyzing relevant factors of automobile sales. Therefore, combing with the above-mentioned literature, we consider not only the external factors influencing the sales volume of the target product, but also the influence of historical data and holiday factors on the product, and establish a combined forecasting model based on multiple linear regression and ARIMA (2, 2, 1) model, to accurately predict the future sales of retail products.

3. Data Processing

The source of the data in this paper is the 2020 Mathorcup College Mathematical Modeling Challenge. We use Excel to select out the data required for the corresponding questions. First, we filter out the top ten target sub categories of sales from June 1 to October 1, 2019, and then process the data of these 10 target sub-categories in 2019, and summarize the daily data into weekly data. A total of 520 sets of data are collected. Each target sub category summarizes 52 weeks data, including sales volume and inventory, actual price, label price, discount, etc. In addition, some missing sales data or influencing factor data in the target sub category are also found when we sort out the data. Therefore, there are four methods to fill in the data. First, if the index value is smooth, we can use the previous data; second, if the data before and after are available, the average value can be used as the missing data; third, if the two groups are similar, we can replace the missing data in a group with the same value in another group; fourth, we use interpolation method to fit the data. After sorting out the complete data, the data of the first nine months of 2019 can be applied to establish a model for fitting, so as to forecast the sales of the top 10 target sub categories in each month of the three months after October 1, 2019, and then determine a model that can accurately predict the demand of new target products.

4. Multiple Linear Regression

Multiple linear regression is a prediction method by establishing the regression function expression making use of the influence of the independent variable on the dependent variable. Various factors all influence the sales volume of target products. The optimal association of many external factors, can help to forecast the future trend of sales data more accurately. Therefore, on the basis of the selected sales data, inventory, actual price, discount, holidays and other factors data. We forecast the sales volume of the target sub category in the next three

months (13 weeks) of 2019 through external macro factors.

4.1. Correlation Analysis

Before performing multiple linear regression, we first make scatter plots of sales volume, price and inventory, and observe the correlation between influencing factors and sales volume.

From Figure 1, it can be shown that the linear relationship between the actual price and the sales volume is not strong, so we take the logarithm of the actual price as an independent variable. It can be shown from Figure 2 that linear correlation between inventory and sales is obviously positive correlated. On the basis of the fact, the inventory of goods is mostly determined by the sales volume. Normly, when sales volume is better, inventory will increase accordingly.

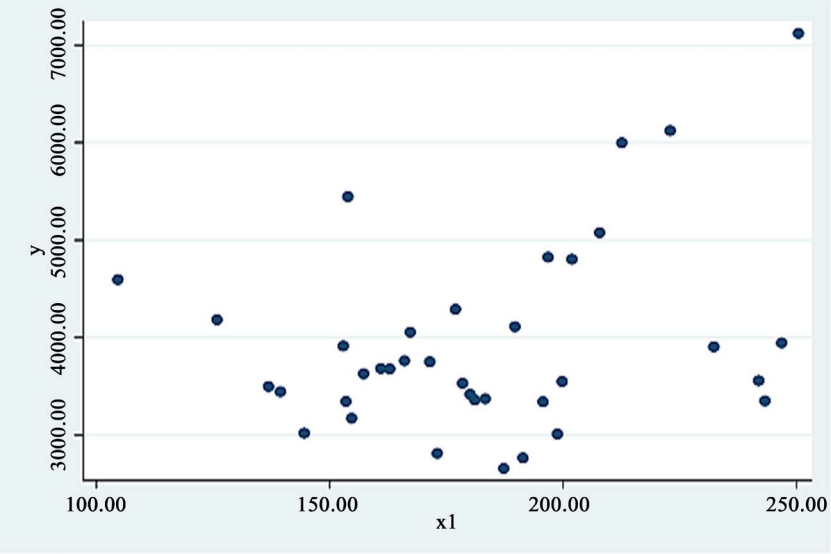

Figure 1. Scatter plot of actual price and sales.

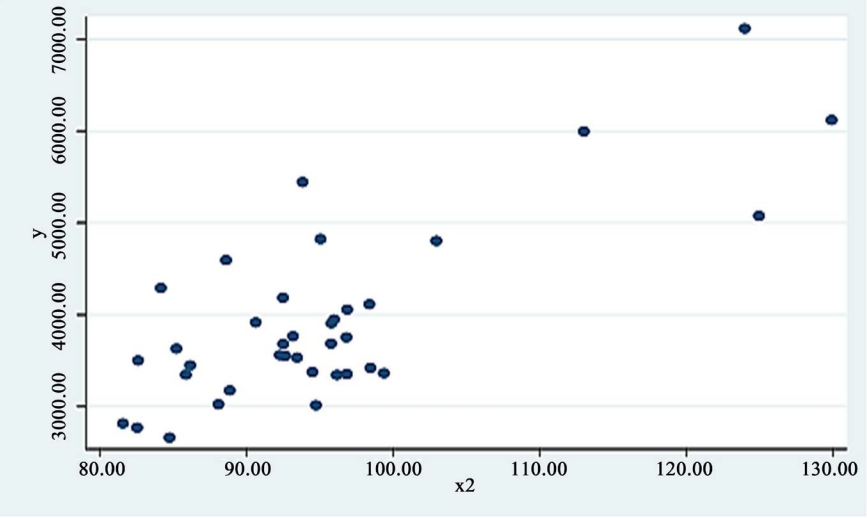

Figure 2. Scatter plot of inventory and sales.

4.2. Model Establishment

We take the sales volume of the target sub category as the dependent variable, and the actual price, inventory, holidays as the independent variables. In fact, holidays are also significantly influence the sales volume of target goods. Generally, before and after New year's day, National day, Double 11 and Double 12, the sales volume of retail enterprises will increase obviously above the normal levels. Therefore, we need to set this factor as a dummy variable. If the week contains holidays, we will take the holiday factor as 1, otherwise we will take it as 0, and establish the following multiple linear regression equation.

$$\begin{cases} y_i = \beta_0 + \beta_1 \ln x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \\ \varepsilon_i \sim N\left(0, \delta^2\right), \ i = 1, \cdots, n \end{cases} \tag{1}$$

The least squares estimation method is used to gauge the parameters, and we make the error sum of squares are smallest.

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 \ln x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3}\right)^2 \tag{2}$$

$$\frac{\partial Q}{\partial \beta_j} = 0, \ j = 0, 1, 2, \cdots, n \tag{3}$$

After sorting out the normal equations, solving the normal equations are as follows

$$\left[\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\right] = \left(X^{\mathrm{T}} X\right)^{-1} X^{\mathrm{T}} Y \tag{4}$$

4.3. Model Solution and Verification

4.3.1. Solving the Model

We use the collected and filtered data for the first 39 weeks of 2019 to solve the multiple linear regression model. In Table 1 and Table 2, the consequences are obtained by using Stata software.

From Table 2, it can be shown:

$$\hat{\beta}_0 = 1588.977, \ \hat{\beta}_1 = -923.520, \ \hat{\beta}_2 = 72.576, \ \hat{\beta}_3 = 773.444$$

Table 1. Multiple linear regression results table.

F	Prob > F	R-squared	Adj R-squared
35.16	0.0000	0.7617	0.7401

Table 2. Multiple linear regression coefficient table.

y	Coef	Std. Err	t	P > |t|
lnprice	−923.520	502.731	−1.87	0.035
inventory	72.576	8.776	8.27	0.000
Holidays	773.444	181.074	4.27	0.000
constant	1588.977	2283.289	0.70	0.491

Then the multiple linear regression model is

$$\hat{y}_i = 1588.977 - 923.520 \ln x_{i1} + 72.576 x_{i2} + 773.444 x_{i3} \quad (5)$$

4.3.2. Model Verification

The hypothesis test of the model is as follows:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0; H_1 : \beta_1, \beta_2, \beta_3$$

$$f = \frac{R^2 / df_e}{\left(1 - R^2\right) / df_r} \sim F\left(df_e, df_r\right)$$

We can see from Table 1 that the P value of the F significance test for the model population is less than 0.05, so we can refuse the original hypothesis, and we can know the overall significance of the model is strong and the overall explanatory ability of the influencing factors to the sales volume is good.

Hypothesis testing of regression coefficients is as follows:

$$H_0 : \beta_j = 0; H_1 : \beta_j \neq 0, \ j = 1, 2, 3$$

$$T = \frac{\hat{\beta}_j - \beta_j}{Se\left(\hat{\beta}_j\right)} \sim t\left(df\right)$$

It can be shown from Table 2 that the P values for the respective T statistics are all less than 0.05, that is, we can refuse the original hypothesis and each influence factor has a good explanation for the sales volume.

4.4. Prediction of Model

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}, 0 \leq R^2 \leq 1$$

For a model, a large coefficient of determination usually corresponds to a high fitting degree. From Table 1, we can observe that the decision coefficient of the model is 0.7617, so we can observe that the predicted value of this model is close to the real value, and the prediction effect is ideal. The model can realize the accurate demand forecast of the target small category products.

5. Establishment and Test of Arima Model

Among the time series models, the ARIMA model is more commonly used. It only needs to use internal previous data and does not need other exogenous variables. The model is denoted as ARIMA (p, d, q), where p is the autoregressive parameter, d is the number of differences required to transform the original non-stationary series into a stationary series, and q is the moving average parameter. Its main modeling steps are shown in Figure 3.

5.1. Stationarity Test and Transformation

Since the establishment of the ARIMA model needs to ensure that the sequence is stable, we use Eviews software to make a sequence diagram for the sales of the

target subcategory in the first 39 weeks of 2019. In order to detect stationarity of the sales volume sequence, the consequences can be observed from the following figure.

It can be seen from Figure 4 that in the first 8 weeks, around the Spring Festival, sales were relatively high, while sales were usually low, showing a seasonal trend. As can be shown from Figure 4 that in the first 8 weeks, around the Spring Festival, sales were relatively high, showing a seasonal trend. On the basis of the fact, the sales volume of the target sub-categories of new retail enterprises will increase in a few days around the holidays. Therefore, we can know that the sales volume sequence is non-stationary.

ADF test is also a widely used method to examine the stability. The existence of unit root is the standard to judge whether the sequence is stable or not. Generally, if the unit root does not exist, the sequence can be judged to be stable, otherwise, it is not stable. This is because when the unit root exists, the regression

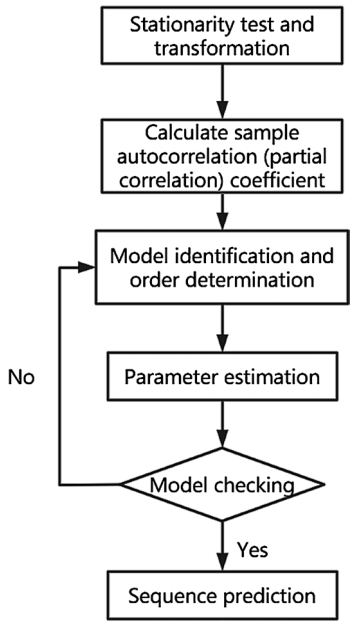

Figure 3. Flow chart of ARIMA model steps.

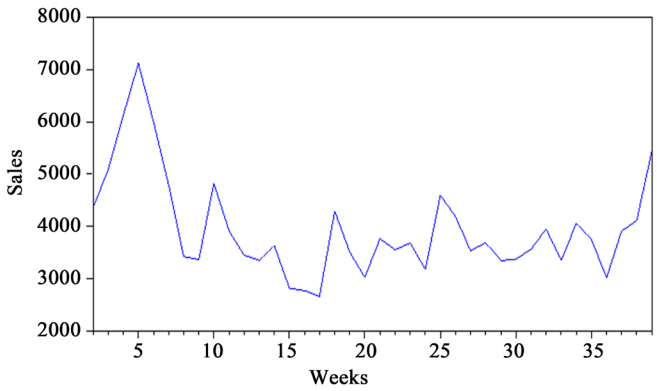

Figure 4. Sequence diagram of sales volume.

is pseudo regression, that is, the error of residual sequence will not decrease with the increase of sample size. Therefore, apart from the timing diagram, ADF test method is also used to further judge the stationarity of the sequence. From the following table, we can get the results.

As can be shown from Table 3 that the p value of ADF test of the sequence is 0.1448. The P value is more than 0.05, so we can accept the original hypothesis. Similarly, we can also know that the sequence is not stable, which corresponds to the image result.

Only a stationary time series can meet the modeling requirements of the ARIMA model, so we need to perform a difference transformation on the non-stationary series.

The ARIMA model is

$$y_t' = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}' + \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} \tag{6}$$

$$y_t' = \Delta^d y_t = (1-L)^d y_t$$

The ARIMA difference model is

$$\left(1 - \sum_{i=1}^{p} \alpha_i L^i\right)(1-L)^d y_t = \alpha_0 + \left(1 + \sum_{i=1}^{q} \beta_i L^i\right)\varepsilon_t \tag{7}$$

We use the Eviews software to perform the first-order difference on the original sequence, and find that the transformed sequence is still not stable, and then perform the second-order difference on it. From Table 4, it can be shown that the t statistic after the second-order difference is −5.746, which is less than −3.646. And it corresponds to a probability is less than 0.05. Therefore, we can observe that the sequence after the second-order difference has passed the ADF test and it is a stationary time series, that is, the value of the difference times d is determined to be 2.

Table 3. ADF test table for sales volume.

		t-Statistic	Pro.*
Augmented Dickey-Fuller-Fuller test statistic		−2.414323	0.1448
Test critical values	1% level	−3.621023	
	5% level	−2.943427	
	10% level	−2.610263	

Table 4. ADF inspection table after second-order difference.

		t-Statistic	Pro.*
Augmented Dickey-Fuller-Fuller test statistic		−5.746084	0.0000
Test critical values	1% level	−3.646342	
	5% level	−2.954021	
	10% level	−2.615817	

5.2. Model Identification and Order Determination

The stationary sales series data processed by the second-order difference has reached the modeling requirements of the ARIMA model. Then, we use Eviews software to make the autocorrelation graph ACF and partial autocorrelation graph PACF of the sales series and determine the value of parameters p and q by the correlation characteristics of the graphs.

We can observe from Figure 5 that the autocorrelation graph of this sequence lags first-order truncation, and the partial autocorrelation graph lags second-order tailing, so the model can be preliminarily determined to be ARIMA (2, 2, 1).

5.3. Model Parameter Estimation

Because there is a little error between autocorrelation graph and partial autocorrelation graph in determining model parameters, sometimes they can not be determined completely and accurately, we compare ARIMA (2, 2, 1) with ARIMA (1, 2, 1) and ARIMA (1, 2, 2) to determine the optimal order and establish a model with the highest fitting degree.

We use Spss to analyze the fitting degree of the three models, and we can get the consequences from Table 5.

From Table 5, we can see that ARIMA (2, 2, 1) has the largest stationary R-square, the largest significance value, and the smallest standard BIC. Therefore, from the point of view of comprehensive indicators, it is obvious that the fit of the ARIMA (2, 2, 1) model is the highest. Therefore, we estimate the parameters

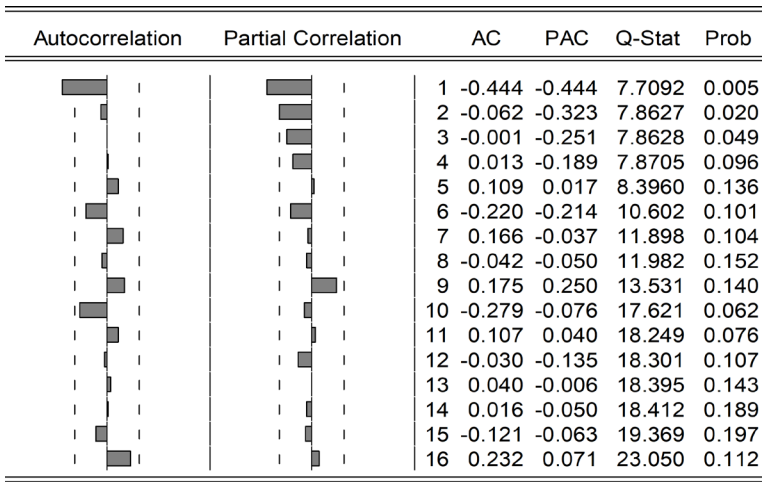

Figure 5. ACF diagram and PACF diagram of the sequence.

Table 5. Fitting statistics of ARIMA model with different parameters.

Model	Stable R Square	Standard BIC	Significance
ARIMA (2, 2, 1)	0.449	13.910	0.357
ARIMA (1, 2, 1)	0.402	13.959	0.316
ARIMA (1, 2, 2)	0.439	13.977	0.288

p = 2, d = 2, q = 1, and establish an ARIMA (2, 2, 1) model.

5.4. Model Test

We can judge whether the residual is a white noise sequence and whether the ARMIA model can well identify the sales volume data by observing the correlation characteristics of the autocorrelation graph and partial autocorrelation graph of the residual.

We can see from Figure 6 that the autocorrelation coefficients and partial autocorrelation coefficients of all lag orders are around 0 and within the range of 2 times the standard deviation. Therefore, we can believe that the residuals are independent, and they are white noise sequence without obvious autocorrelation. The model can recognize the sales volume data very well.

5.5. Prediction of the Model

We compare the actual sales volume in the first 9 months of 2019 with the sales volume fitted by the ARIMA (2, 2, 1) model. We can observe from Figure 7

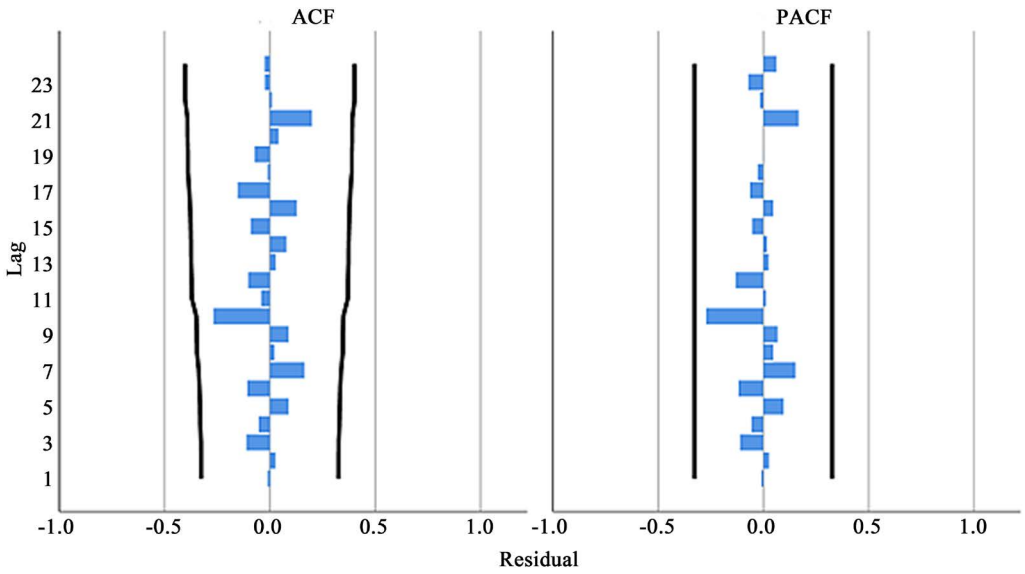

Figure 6. Residual ACF and PACF plots of the ARIMA model with different parameters.

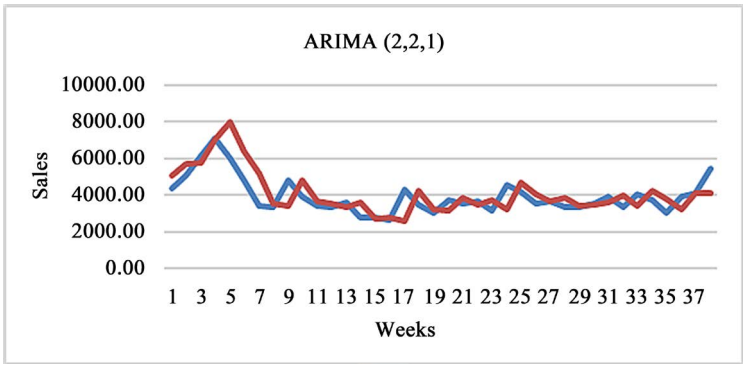

Figure 7. Fitting diagram of ARIMA model with different parameters.

that the change trend of the real sales data and the fitted sales data are roughly the same. Therefore, we can observe that the ARIMA (2, 2, 1) model has a better fitting effect.

6. Establishment of Combination Model

We first establish a multiple linear regression model and consider macroscopic influencing factors when predicting the sales volume of the target subcategory. Secondly, because the time series use the regularity of their own data to predict, the previous sales data will influence the current sales, so we build an ARIMA model, and compare the ARIMA models with different parameters respectively, and finally establish the optimal ARIMA (2, 2, 1) model. Analyzing these two models, time series analysis can find trends and seasonal factors, such as the holiday factors and the internal law of one's own data can be fully utilized. But time series analysis does not consider macroscopic factors. Multiple linear regression thinks over macroscopic factors, but it cannot use the trend and seasonal characteristics of the data and if the two change, it can't cope well. Therefore, which prediction method is used alone is relatively one-sided. However, a combined model that makes full use of the advantages of the two models will make the prediction results more accurate and more robust.

Synthesize the above research, we integrate the multiple linear regression model with ARIMA (2, 2, 1) model, and assign different weights to the prediction values of the two single models on the basis of the degree of fit, and further predict the true values. We take $\hat{y}_1$ as the predicted value of multiple linear regression and $\hat{y}_2$ as the predicted value of the ARIMA (2, 2, 1) model. Then, in view of their respective degrees of fit, a weight of 0.4 is assigned to $\hat{y}_1$, and a weight of 0.6 is assigned to $\hat{y}_2$, then we can build a combination model.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{100\left|y_i - \hat{y}_i\right|}{y_i} \tag{8}$$

Finally, we use formula (8) to calculate the average MAPE values of the multiple linear regression model, the ARIMA (2, 2, 1) model, and the combined model to be 13.462, 11.826, 9.437, respectively. As can be shown that the combined model not only considers the impact of internal preliminary data, but also considers external factors, so that the forecast accuracy is further improved, and the needs prediction for the target goods of the new retail enterprise is more accurate.

7. Conclusion

In the period of the new retail era, consumer experience and needs are the most significant aspects for an enterprise to be concerned with. In order to satisfy the decentralized and differentiated needs of consumers, enterprises need to provide consumers with more kinds of goods, which also needs enterprises to have excellent abilities to manage inventory and formulate reasonable and effective

production plans, so that the goods provided can meet the needs of consumers without causing inventory accumulation and waste of resources. The precise forecast of the demand for retail goods with complex levels and various varieties will be the prerequisite for enterprises to make reasonable decisions. Therefore, on the one hand, this article builds a multiple linear regression model to research the forecast of actual price, inventory, and holiday on sales. On the other hand, we utilize the characteristics of the historical data tendency of the target goods, through parameter estimation and fitting degree comparison, to establish the optimal ARIMA (2, 2, 1) model. Since a single prediction is difficult to accurately predict the target product with complex levels, we finally combine the advantages of the two models to establish a combined prediction model on the basis of multiple linear regression and ARIMA (2, 2, 1). We can get the consequences that the MAPE value of the combined model is 9.437, the prediction effect is better. It can achieve precise forecast of various target goods at distinct levels, and help business leaders make scientific and effective management decisions, thereby reducing the difficulty of inventory management, reducing capital occupation, increasing economic benefits, meeting consumer demand, and enhance the brand influence of enterprises, enhance their competitiveness, and promote the further development of new retail enterprises.

The combined forecasting model based on multiple linear regression and ARIMA established in this paper only studies the linear characteristics of the target product sales data. In the future, a new combination forecasting model can be established on this basis to further study the nonlinear characteristics of the sales data to obtain more accurate prediction results.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Dong, T. T., Dong, X. S., Zhang, R., & Cui, J. F. (2020). Enterprise Sales Forecast Based on Exponentially Weighted Moving Average Method. *Journal of Qingdao University (Natural Science Edition), 33,* 50-54.

Gong, W. W., & Huang, J. (2017). Comprehensive Model of Demand Forecast Based on Grey Theory and Exponential Smoothing Method. *Statistics and Decision, No. 1,* 72-76.

Liang, Y. D. (2018). *Research on Hotel Online Sales Forecast Based on Combination Model.* Xi'an: Xidian University.

Miao, H., Tang, C. T., & Luo, L. L. (2020). New Energy Vehicle Sales Forecast Based on ARIMA Model. *Enterprise Technology and Development, No. 10,* 97-98.

Rong, F. Q., & Guo, M. F. (2019). Research on Online Product Sales Forecast Analysis Based on Convolutional Neuralnetwork. *Journal of Northwest University for Nationalities (Philosophy and Social Sciences Edition), No. 2,* 15-26.

Wang, Y. (2019). *Research on the Status Quo and Forecast of My Country's Heavy Truck*

Sales Based on Factor Analysis. Jinan: Shandong University.

Wu, M., Lin, H. P., Li, S. K., Wu, M. Z., Wang, Z. G., & Wu, G. F. (2016). A Prediction Method of Cigarette Sales Based on Support Vector Machine. *Tobacco Science and Technology, 49,* 87-91.

Yang, B. R. (2017). *Passenger Car Market Prediction Model Based on Multiple Linear Regression and BP Neural Network.* Wuhan: Huazhong University of Science and Technology.

Zhang, C., & Qiu, T. (2019). Gas Station Sales Forecast Based on Decision Tree Integration Model. *Computers and Applied Chemistry, 36,* 615-619.

Zhang, J. R. (2020). *Research on Combined Forecasting Model of Dish Sales Based on Time Series and Neural Network.* Hangzhou: Hangzhou Dianzi University.