

# ***Time Series Analysis and Forecast of Sales of New Car and Used Car Using SARIMA Model***

**Junkai Chen<sup>1,a,\*</sup>**

<sup>1</sup>*College of Economics, Shenzhen University, Taoyuan District, Shenzhen, China*  
*a. chenjunkai2021@email.szu.edu.cn*

*\*corresponding author*

**Abstract:** This paper conducts a comprehensive time series analysis of new and used car sales in the United States, focusing on intrinsic patterns captured by Seasonal AutoRegressive Integrated Moving Average (SARIMA) models. SARIMA models are applied to forecast sales over the next two years, and select the model based on standards such as Akaike Information Criterion with correction (AICc), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Notably, the ARIMA (4, 0, 3) (3, 1, 1)<sup>[12]</sup> model emerges as the optimal fit for new car sales, displaying superior time series fitting and lower errors. For used car sales, the ARIMA (2, 0, 3) (2, 1, 3)<sup>[12]</sup> model, although not the best-fitting, exhibits the lowest prediction errors. Consequently, these models are chosen for forecasting. The results suggest a continued upward trajectory in new and used car sales in the United States over the next two years, capturing the inherent cyclic and seasonal patterns inherent in the data.

**Keywords:** SARIMA model, AICc, RMSE, MAE, forecast

## **1. Introduction**

In contemporary society, automobiles stand as one of the most integral and convenient modes of transportation, often chosen by many for their travel needs. The analysis and prediction of new and used car sales emerge as pivotal elements, carrying substantial implications across diverse economic landscapes. Such insights prove invaluable to both suppliers and consumers, fostering a more profound comprehension of the automotive sales market dynamics.

This research delves into the application of SARIMA models, a promising avenue for understanding and forecasting sales trends. By employing SARIMA models on two distinct time series, this study aims to predict future sales, shedding light on the accuracy of short-term forecasts. This dual-pronged approach not only facilitates sellers in formulating effective sales strategies but also empowers buyers to make informed purchasing decisions.

An integral component of contemporary business intelligence is sales forecasting [1]. Sales forecasting is recognized as a complex challenge, wherein accurate predictive models assist businesses in uncovering potential risks and making more informed decisions [2,3]. Forecasts hold significance in the automotive industry, as precise predictions can effectively reduce inventory buildup, prevent missed sales opportunities, and mitigate issues of over-supply [4,5].

Simultaneously, several studies have successfully employed SARIMA models for real-time series analysis and forecasting [6-10]. These studies underscore the significance of sales forecasting and the feasibility of SARIMA models for real-time series prediction.

This paper conducts a comprehensive time series analysis of new and used car sales in the United States, selecting SARIMA models based on standards such as AICc, RMSE, MAE, and MAPE. The subsequent forecasting of sales for the upcoming two years aims to provide a robust understanding of the market dynamics and trends, contributing to both academic discourse and practical applications within the automotive industry.

## 2. Methodology

### 2.1. Description of Data

The dataset used in this study is sourced from the United States Census Bureau (<http://www.census.gov/>), consisting of monthly figures for new car sales and used car sales. The data spans from January 1992 to November 2023 and is presented in Figure 1. For new car sales, the trend shows fluctuating increases, but experienced significant declines around 2009 and 2022, respectively. Similarly, for used car sales, the overall trend is characterized by fluctuating increases, with substantial declines in the same years. However, compared to the sales of new car, the monthly sales and volatility are lower for used car.

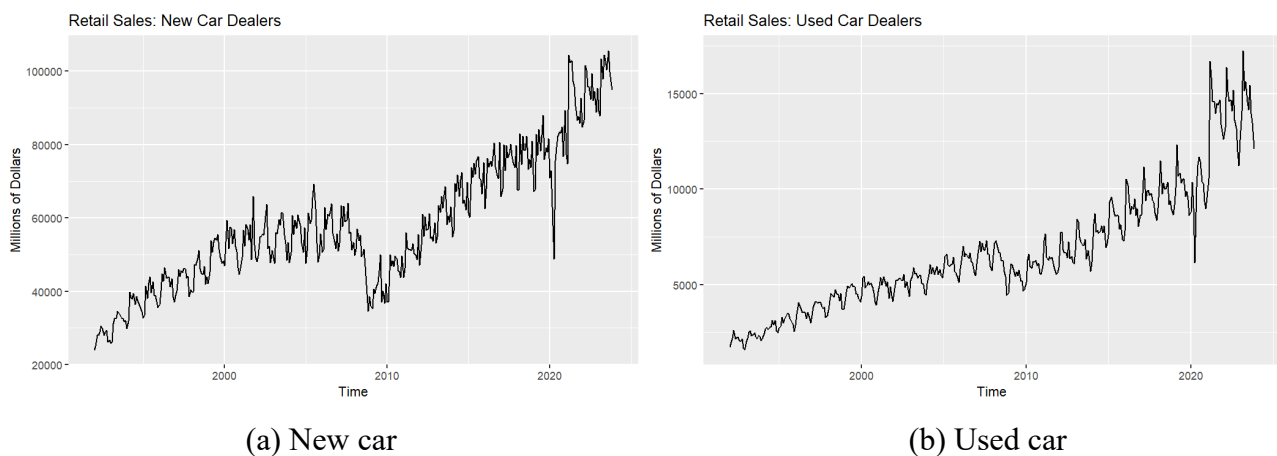
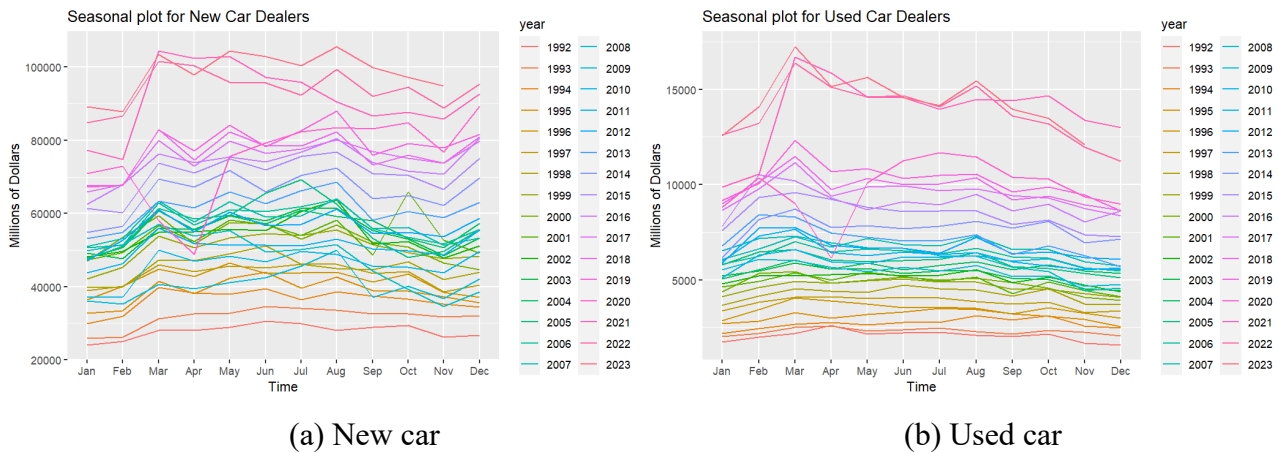


Figure 1: Retail sales for new car(a) and retail sales for used car (b).

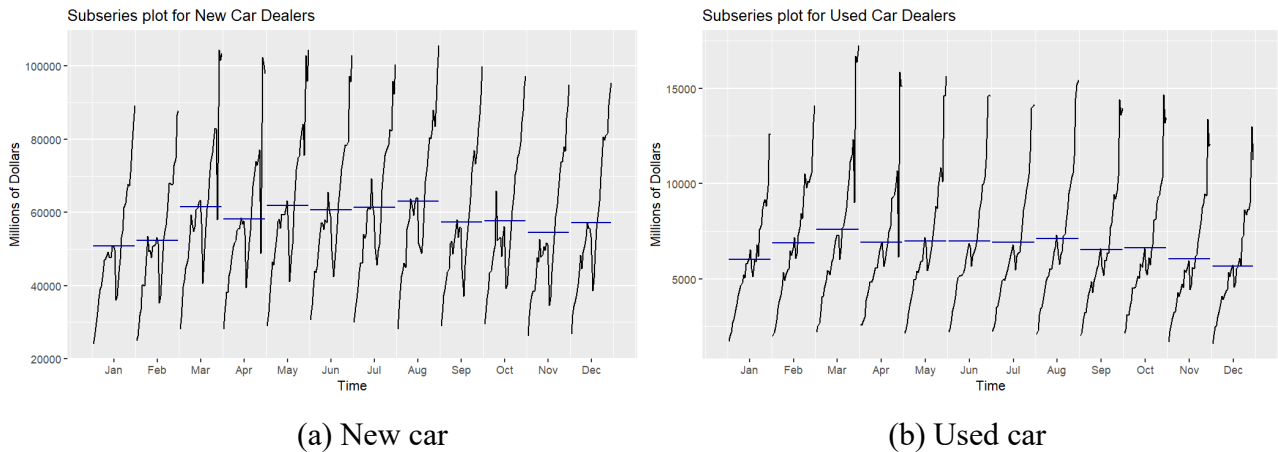
### 2.2. Time Series Patterns Analysis

To better fit models later, now conduct an analysis of these time series. From Figure 2, it can be observed that for new car, retail sales increase slightly in February, then fluctuates. For used car, retail sales rise in February and March, then slowly declines.



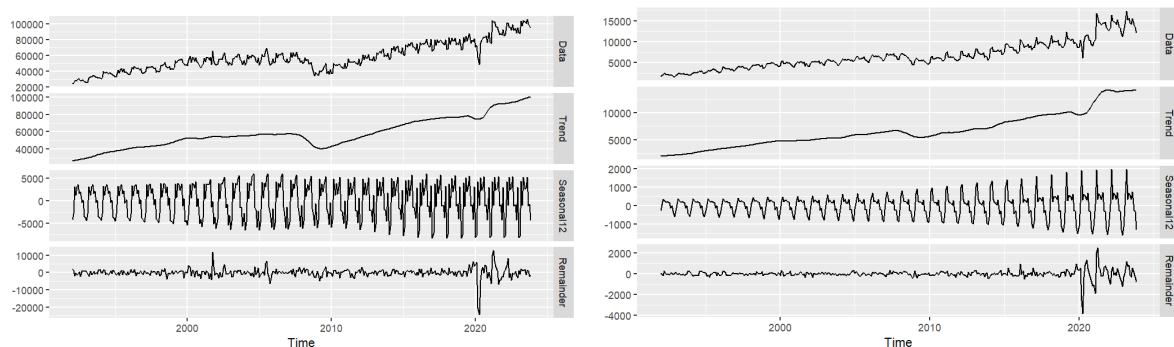
(a) New car (b) Used car  
Figure 2: Seasonal plot for new car sales(a) and used car sales(b).

Figure 3 further illustrates the trends in the data, and it is evident that the trends of these time series are clearly visible.



(a) New car (b) Used car  
Figure 3: Subseries plot for new car sales(a) and used car sales(b).

To further identify the components contained in the time series, a decomposition of these time series was conducted, and the results are illustrated in Figure 4. It can be noted that these time series demonstrate similar trends. In general, sales gradually increase over time, with slight decreases and fluctuations around 2009 and 2021. Additionally, both time series show seasonality, and the seasonality in the sales of new car is more pronounced, suggesting the potential need for a seasonal ARIMA model.



(a) New car (b) Used car  
Figure 4: The components of new car sales(a) and used car sales(b).

### 2.3. Seasonal ARIMA

The ARIMA model necessitates the precondition of data stationarity. Prior to model fitting, the Augmented Dickey-Fuller (ADF) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) tests are executed to evaluate the stationarity of the time series. The ADF test is based on the assumption that, under the null hypothesis, the data is non-stationary, while it is considered stationary under the alternative hypothesis. In contrast, the KPSS test posits that, under the null hypothesis, the data is stationary, but under the alternative hypothesis, it is non-stationary. The outcomes of the tests are presented in Table 1, revealing that both time series demonstrate non-stationarity based on the p-values.

Table 1: The results of the ADF test and KPSS test (Raw data).

	New car sales	Used car sales
p-value of ADF test	0.5417	0.1968
p-value of KPSS test	<0.01	<0.01

After applying first-order seasonal differencing and logarithmic transformation to the time series, the outcomes of the tests are presented in Table 2, revealing that both time series have become stationary based on the p-values.

Table 2: The results of the ADF test and KPSS test (Processed data).

	New car sales	Used car sales
p-value of ADF test	<0.01	0.01502
KPSS test	>0.1	>0.1

After applying first-order seasonal differencing and logarithmic transformation to the time series, the time series plots are presented in Figure 5.

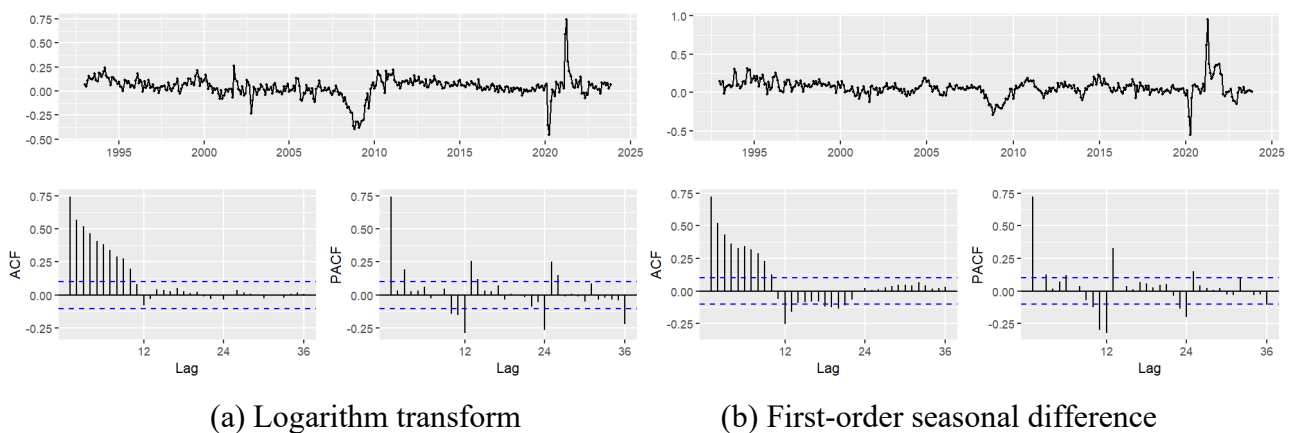


Figure 5: Time series plots for data (Logarithm transform and first-order seasonal difference).

Meanwhile, the Partial AutoCorrelation Function (PACF) plots in Figure 5 reveal a strong seasonality in both time series. Therefore, utilizing a Seasonal ARIMA model for fitting would better capture the characteristics of the time series.

The SARIMA model is a powerful instrument in time series analysis for forecasting and modeling. It is an extension of the Autoregressive Integrated Moving Average (ARIMA) model specifically designed to handle time series data with seasonal variations. The SARIMA model's principles encompass four key components: Seasonal Autoregressive, Seasonal Differencing, Seasonal Moving

Average, and the three components of the non-seasonal ARIMA model: Autoregressive, Differencing, and Moving Average [11].

Equation (1) can represent the SARIMA (p, d, q) (P, D, Q)<sup>[m]</sup> model:

$$\phi_p(B)\Phi_P(B^m)(1-B)^d(1-B^m)^D y_t = \theta_q(B)\Theta_Q(B^m)\varepsilon_t \quad (1)$$

Where  $p$  denotes the magnitude of non-seasonal AR term,  $P$  denotes the magnitude of seasonal AR term,  $q$  denotes the magnitude of non-seasonal MA term,  $Q$  denotes the magnitude of seasonal MA term,  $d$  represents the magnitude of non-seasonal differencing,  $D$  denotes the magnitude of seasonal differencing,  $m$  denotes the number of observations per year,  $y_t$  denotes the observed values.  $B$  represents the backshift operator which can be denoted as the following equation 2:

$$B^m y_t = y_{t-m} \quad (2)$$

### 3. Empirical Results

For new car sales, from the previous analysis, it was determined that due to the non-stationarity of the data and the significant seasonality, first-order seasonal differencing is required. Therefore, it can be preliminarily established that  $D=1$  and  $d=0$ . The AutoCorrelation Function (ACF) plot and PACF plot of this time series after logarithmic transformation and first-order seasonal differencing are illustrated in Figure 6. The notable spike at lag 1 in the ACF plot implies the presence of a non-seasonal MA (1) component, while the absence of a significant spike at lag 12 in the ACF plot, and similarly at lag 24, leads to the inference that the magnitude of the seasonal MA component is considered as 0. Meanwhile the notable spike at lag 3 in the PACF plot suggests a non-seasonal AR (3) component, and the notable spike at lag 12 and 24 in the PACF plot indicates the presence of a seasonal AR (2) component.

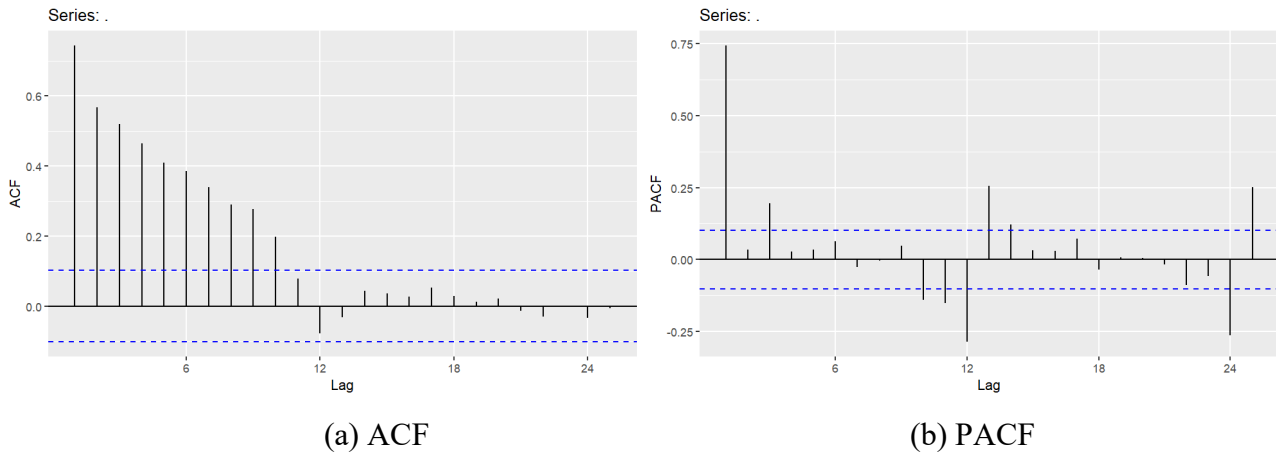


Figure 6: ACF plot and PACF plot for new car sales (Logarithm transform and first-order seasonal difference).

Therefore, the model is initially set as ARIMA (3, 0, 1) (2, 1, 0)<sup>[12]</sup>. Subsequently, multiple models are fitted, and the fitted models along with their corresponding AICc values, RMSE, MAE, and MAPE are presented in Table 3. By comparing these values, a better-fitting model can be selected. According to the AIC criterion, a superior model has a lower AICc value. Similarly, a model with higher accuracy should have lower errors.

Table 3: Postulated models and corresponding value.

Model	AICc	RMSE	MAE	MAPE
ARIMA (3, 0, 1) (2, 1, 0) <sup>[12]</sup>	-1000.42	0.0601	0.0423	0.3861
ARIMA (3, 0, 1) (2, 1, 1) <sup>[12]</sup>	-1051.92	0.0554	0.0380	0.3477
ARIMA (3, 0, 2) (2, 1, 1) <sup>[12]</sup>	-1054.57	0.0550	0.0377	0.3452
ARIMA (3, 0, 2) (2, 1, 2) <sup>[12]</sup>	-1054.71	0.0548	0.0374	0.3426
ARIMA (3, 0, 3) (3, 1, 1) <sup>[12]</sup>	-1057.87	0.0545	0.0370	0.3382
ARIMA (4, 0, 3) (3, 1, 1) <sup>[12]</sup>	-1066.53	0.0536	0.0357	0.3265

By comparing AICc values, RMSE, MAE, and MAPE, it is observed that the ARIMA (4, 0, 3) (3, 1, 1) <sup>[12]</sup> model is the best-fitting model and also exhibits the highest accuracy. It is noted that, in this case, models with better fit tend to have higher accuracy.

Subsequently, a residual analysis is conducted, and Figure 7 illustrates the results. Furthermore, the Ljung-Box test results in a p-value of 0.2832, implying the acceptance of the null hypothesis and indicating that the residuals show no autocorrelation.

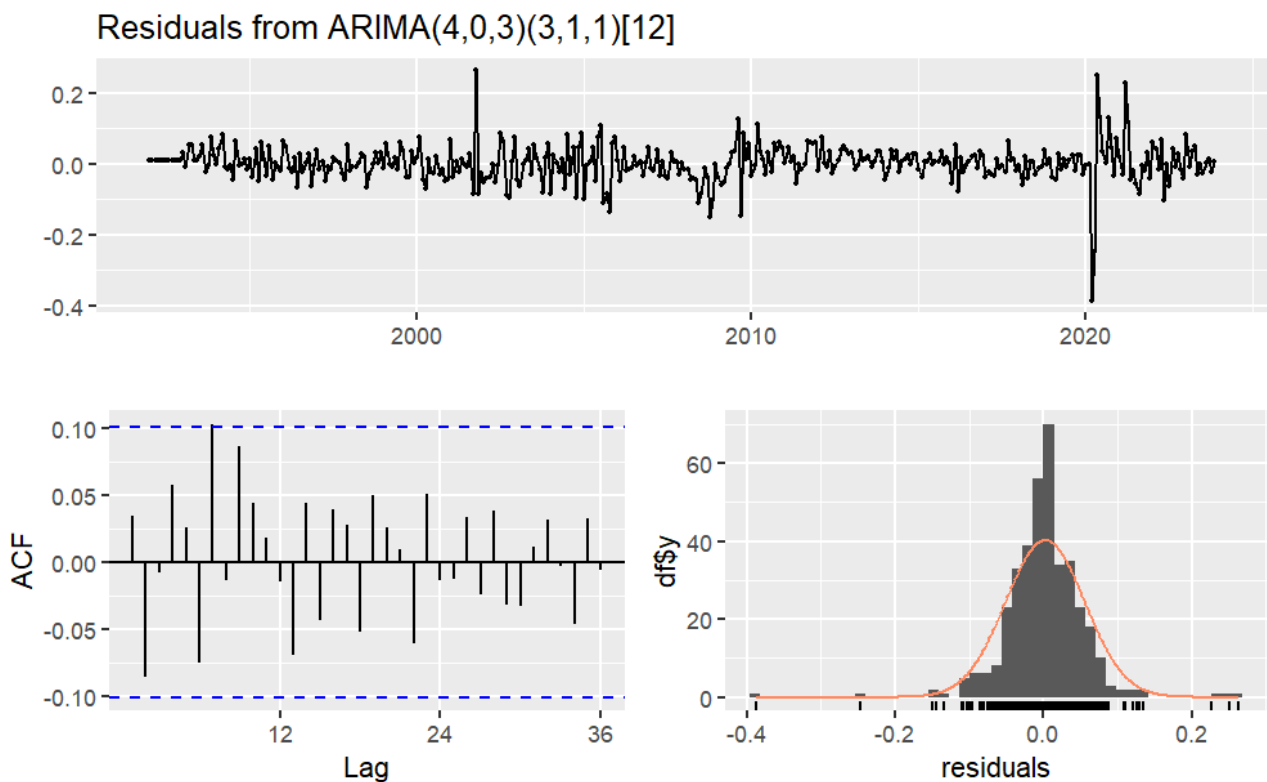


Figure 7: Residuals plot from ARIMA (4,0,3) (3,1,3) <sup>[12]</sup>

Table 4 presents detailed information on the parameters of the ARIMA (4,0,3) (3,1,3) <sup>[12]</sup> model. From the table, it can be observed that, for the sales of new cars, lag periods 1, 2, and 3 exhibit positive autocorrelation with the current period, while lag period 4 shows negative autocorrelation with the current period. Regarding the seasonal lag periods, 1 to 3 exhibit negative autocorrelation with the current period, indicating that an increase in sales in the previous three periods corresponds to a concurrent increase in the current period. From the perspective of seasonal autocorrelation, an increase in sales in the same period of the previous year results in a decrease in the current period's sales.

Analyzing the coefficients of the MA components, the coefficient for MA(1) is significant and positive, indicating that the current observation is positively influenced by past errors. Similarly, the coefficient for SMA(1) is significant and negative, signifying that the current observation is negatively impacted by seasonal errors.

Table 4: Coefficients of ARIMA (4,0,3) (3,1,3) <sup>[12]</sup>.

Components	Coefficient	Standard Error
AR(1)	0.2800	0.0986
AR(2)	0.2153	0.0157
AR(3)	0.9263	0.0152
AR(4)	-0.4298	0.0981
MA(1)	0.4189	0.0768
MA(2)	0.1425	0.0741
MA(3)	-0.7025	0.0780
SAR(1)	-0.1251	0.1113
SAR(2)	-0.1600	0.0931
SAR(3)	-0.0408	0.0871
SMA(1)	-0.7046	0.0971

Next, the model is used to forecast the sales volume of new cars for the next two years. As depicted in Figure 8, the forecasted results indicate that sales will continue to exhibit a trend of fluctuating upward growth over the next two years.

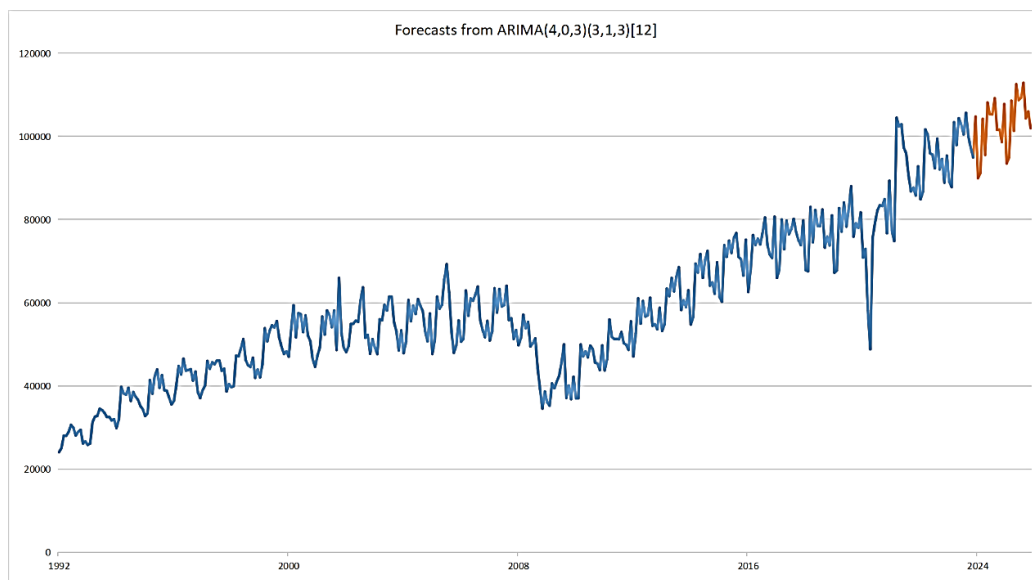


Figure 8: Forecasts from ARIMA (4,0,3) (3,1,3) <sup>[12]</sup>

Similarly, for used car sales, from the previous analysis, it was determined that due to the non-stationarity of the time series and the significant seasonality, first-order seasonal differencing is required. Therefore, it can be preliminarily established that  $D=1$  and  $d=0$ . The ACF plot and PACF plot of this time series after logarithmic transformation and first-order seasonal differencing are illustrated in Figure 9. The notable spike at lag 1 in the ACF plot implies the presence of a non-seasonal MA (1) component, while the notable spike at lag 12 in the ACF plot leads to the inference that the magnitude of the seasonal MA component is considered as 1. Meanwhile the notable spike

at lag 2 in the PACF plot implies the presence of a non-seasonal AR (2) component, and the notable spike at lag 12 and 24 in the PACF plot indicates the presence of a seasonal AR (2) component.

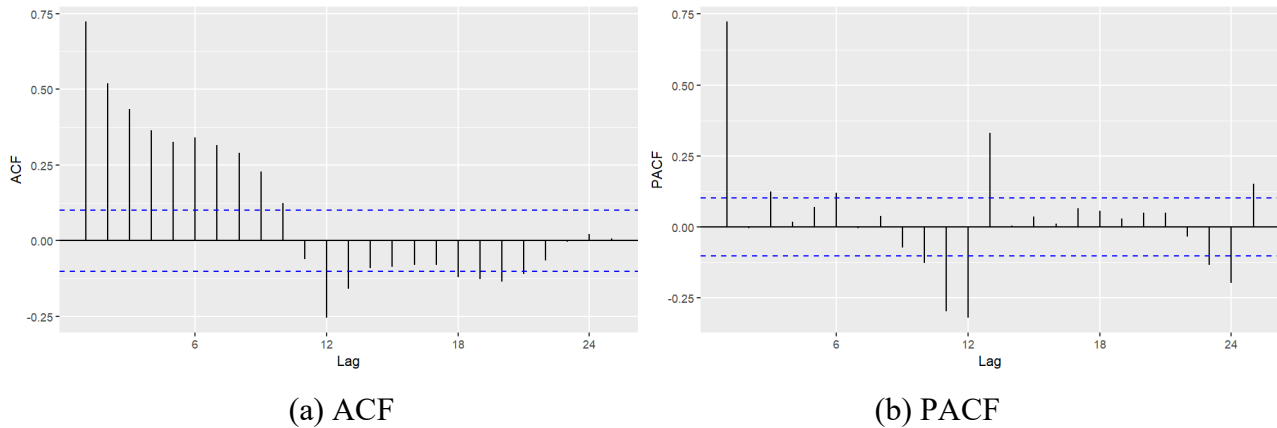


Figure 9: ACF plot and PACF plot for used car sales (Logarithm transform and first-order seasonal difference).

Therefore, the model is initially set as  $ARIMA(2, 0, 1)(2, 1, 1)^{[12]}$ . Subsequently, multiple models are fitted, and the fitted models along with their corresponding AICc values, RMSE, MAE, and MAPE are presented in Table 5. Continue to identify a better-fitting model by comparing these values.

Table 5: Postulated models and corresponding value.

Model	AICc value	RMSE	MAE	MAPE
$ARIMA(2, 0, 1)(2, 1, 1)^{[12]}$	-1003.91	0.0594	0.0400	0.4583
$ARIMA(2, 0, 2)(2, 1, 1)^{[12]}$	-1001.99	0.0595	0.0401	0.4598
$ARIMA(2, 0, 3)(2, 1, 3)^{[12]}$	-1001.77	0.0589	0.0398	0.4555
$ARIMA(2, 0, 1)(2, 1, 2)^{[12]}$	-1002.19	0.0593	0.0399	0.4570

By comparing AICc values, it was found that  $ARIMA(2, 0, 1)(2, 1, 1)^{[12]}$  is the best-fitting model among these models. However, the  $ARIMA(2, 0, 3)(2, 1, 3)^{[12]}$  model has the lowest RMSE, MAE, and MAPE. In the pursuit of higher forecast accuracy, the final choice is to use the  $ARIMA(2, 0, 3)(2, 1, 3)^{[12]}$  model for forecasting the used car sales. It is noted that, in this case, models with better fit tend to have higher accuracy.

Subsequently, a residual analysis is conducted, and Figure 10 illustrates the results. Furthermore, the Ljung-Box test results in a p-value of 0.2039, implying the acceptance of the null hypothesis and indicating that the residuals show no autocorrelation.



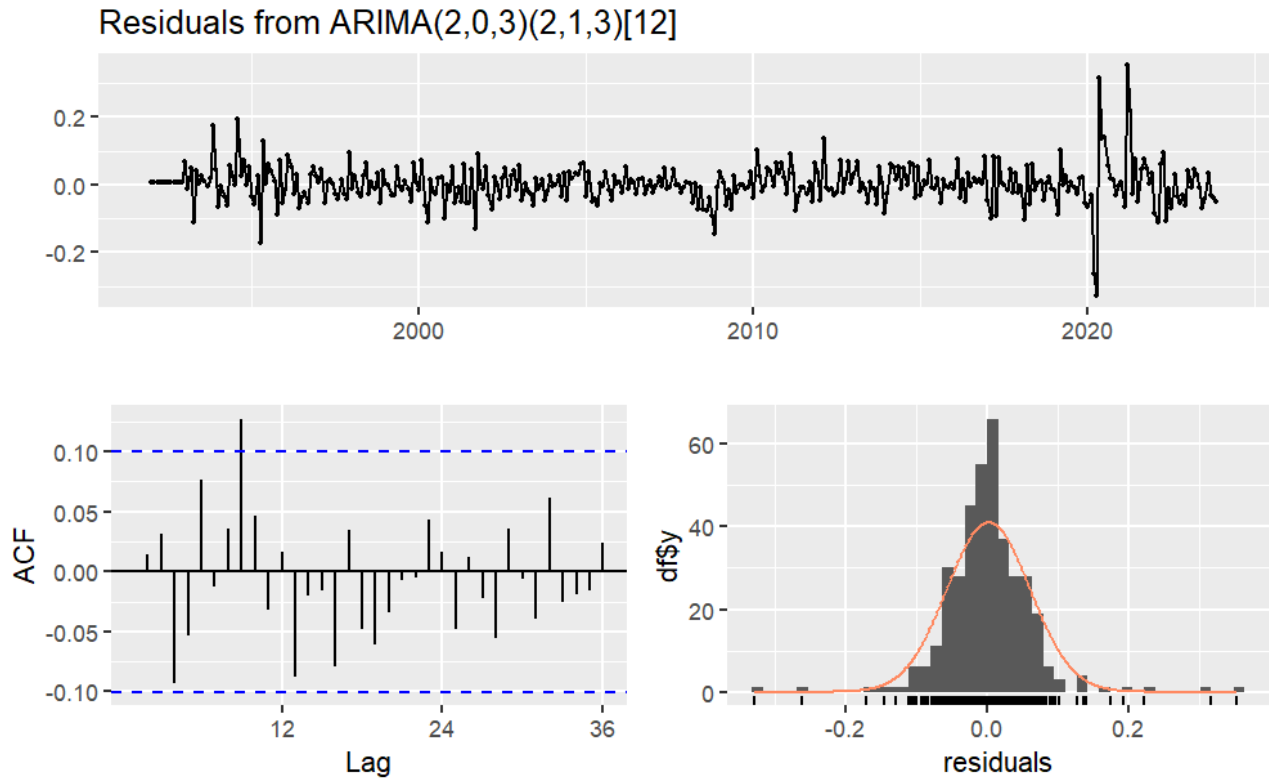


Figure 10: Residuals plot of ARIMA (2,0,3) (2,1,3) <sup>[12]</sup>

Table 6 displays detailed information about the parameters of the ARIMA (2,0,3) (2,1,3) <sup>[12]</sup> model. From the table, it is evident that, for the sales of used cars, lag period 2 exhibits a significant positive autocorrelation with the current period. Regarding the seasonal lag period, the coefficient for SAR(1) is significant and negative, indicating that an increase in sales in lag period 2 corresponds to a concurrent increase in the current period. From the perspective of seasonal autocorrelation, an increase in sales in the same period of the previous year results in a decrease in the current period's sales.

Analyzing the coefficients of the MA components, the coefficient for MA(1) is significant and positive, signifying that the current observation is positively influenced by past errors. Similarly, the coefficients for MA(2) and MA(3) are significant and negative, indicating that the current observation is negatively impacted by past errors. Likewise, the coefficient for SMA(2) is significant and negative, suggesting that the current observation is negatively influenced by seasonal errors.

Table 6: Coefficients of ARIMA (2,0,3) (2,1,3) <sup>[12]</sup>.

Components	Coefficient	Standard Error
AR(1)	0.0521	0.0485
AR(2)	0.9413	0.0486
MA(1)	0.6424	0.0670
MA(2)	-0.5116	0.0580
MA(3)	-0.2642	0.0533
SAR(1)	-0.8168	0.7899
SAR(2)	0.1102	0.6821
SMA(1)	-0.0496	0.7857
SMA(2)	-0.7041	0.1684
SMA(3)	0.1577	0.5049

Next, the model is used to forecast the sales volume of used car for the next two years. As depicted in Figure 11, the forecasted results indicate that sales will continue to exhibit a trend of fluctuating upward growth over the next two years.

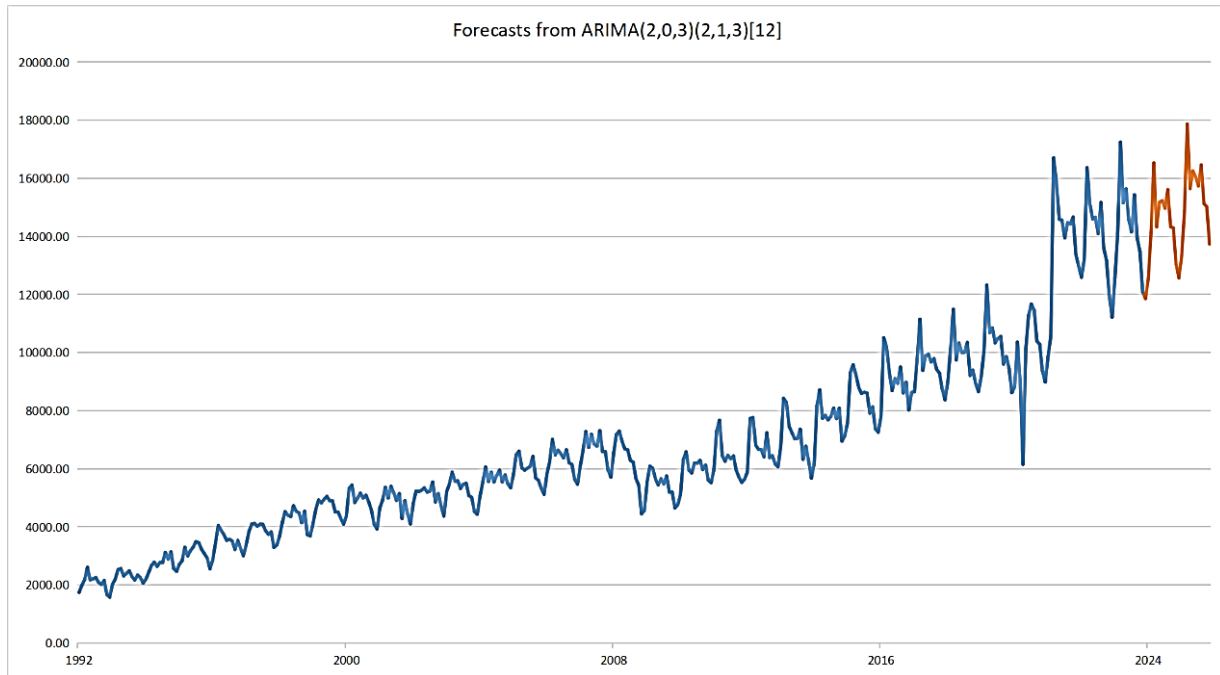


Figure 11: Forecasts from ARIMA (2,0,3) (2,1,3) <sup>[12]</sup>

#### 4. Conclusion

In conclusion, this paper of new and used car sales time series in the United States using SARIMA models has yielded valuable insights into the overall trends and seasonality of the market. While the selected models, ARIMA (4, 0, 3) (3, 1, 1) <sup>[12]</sup> for new car sales and ARIMA (2, 0, 3) (2, 1, 3) <sup>[12]</sup> for used car sales, have demonstrated efficacy in capturing time series characteristics and forecasting, there are notable considerations and opportunities for improvement.

The coefficients of the model also reveal that, for new car sales, the presence of positive autocorrelation in lag periods 1, 2, and 3, along with negative seasonal autocorrelation in lag period 1, suggests that an increase in sales in the previous three periods will lead to a corresponding increase in the current period's sales. From the perspective of seasonal autocorrelation, however, an increase in sales in the same period of the previous year will result in a decrease in the current period's sales. Furthermore, the current period's sales are positively influenced by past errors and negatively impacted by seasonal errors.

On the other hand, for used car sales, the existence of positive autocorrelation in lag period 2 and negative seasonal autocorrelation in lag period 1 indicates that an increase in sales in lag period 2 will correspondingly increase the current period's sales. From the perspective of seasonal autocorrelation, an increase in sales in the same period of the previous year will lead to a decrease in the current period's sales. Additionally, the current period's sales are influenced positively by MA(1) and negatively by MA(2) and MA(3), along with negative impacts from seasonal errors.

The practical implications of this paper extend to both sellers and buyers in the automotive industry, providing a basis for strategic decision-making. However, it is essential to acknowledge the limitations of our study, including the reliance on a specific model type and the exclusion of certain

influencing factors. Future research endeavors could explore incorporating additional models such as VAR or GARCH, offering a more comprehensive understanding of market dynamics.

The challenges in model selection, particularly the discrepancy between AICc values and forecast errors in used car sales, highlight the intricacies of forecasting in dynamic markets. Exploring the reasons behind this divergence and refining the model selection process will be crucial for advancing forecast accuracy.

While the forecasts indicate sustained upward-trending fluctuations in sales volumes over the next two years, it is worth noting the importance of considering external factors that could influence these trends. Decision-makers should approach these forecasts with an awareness of the ever-changing landscape and be prepared to adapt strategies accordingly.

In recommending further avenues for research, it is worth noting the exploration of alternative models and methods, such as artificial neural networks, to enhance forecast accuracy and uncover additional nuances within the time series. The iterative nature of forecasting underscores the need for continuous improvement and adaptation to evolving market conditions, reinforcing the dynamic relevance of our research in the automotive industry.

## References

- [1] Kohli, S., Godwin, G. T., & Urolagin, S. (2020). Sales prediction using linear and KNN regression. In *Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019* (pp. 321-329): Springer.
- [2] Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, 170, 97-135.
- [3] Wenshun, Sheng., Hanchi, Zhao., & Yanwen, Sun. (2019). Sales Forecasting Model Based on Improved Genetic Algorithm Optimized BP Neural Network. *Computer Systems & Applications*, 28(12), 200-204.
- [4] Permatasari, C. I., Sutopo, W., & Hisjam, M. (2018). Sales forecasting newspaper with ARIMA: A case study. Paper presented at the AIP Conference Proceedings.
- [5] Pavlyshenko, B. M. (2019). Machine-learning models for sales time series forecasting. *Data*, 4(1), 15.
- [6] Jingli, Guo., & Bo, Dong. (2019). International Rice Price Forecast Based on SARIMA Model. *Price Theory and Practice*, 01, 79-82.
- [7] Zhilin, Chen., Hongtao, Hu., & Yingying, Bian. (2024). Forecasting Shanghai Port Container Throughput under the COVID-19 Pandemic Using the SARIMA Model. *Industrial Engineering and Management*, 01, 32-40.
- [8] Farhan, J., & Ong, G. P. (2018). Forecasting seasonal container throughput at international ports using SARIMA models. *Maritime Economics & Logistics*, 20, 131-148.
- [9] Hao, Huang., & Yingmei, Deng. (2023). Forecast Analysis of Neurosurgical Workload in a Tertiary Hospital Based on the SARIMA Model. *Chinese Medical Records*, 11, 52-54.
- [10] Wanning, Sun., Jing, Yang., & Yiyi, Yang. (2018). Analysis and Forecast of China's GDP Based on SARIMA Model. *Chinese Collective Economy*, 36, 78-80.
- [11] Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). *Time series analysis and its applications* (Vol. 3). New york: Springer.