Time series analysis of sales data by statistical methods

Abstract

Time series analysis is a common statistical method to analyse time series data which can be found in many fields, such as economics and finance. This report focuses on the statistical methods of time series analysis with an example of a financial time series data. It conducts an exploratory data analysis to suggest possible models; suggested models are then compared by model performance. Based on the best fitted model, future values are predicted; spectral analysis is also performed to study the periodicity of time series. Finally, we discuss the methods and findings.

Keywords: Time series, ARIMA model, ACF, PACF, sales, forecast

Introduction

Time series data is the data collected at different points of time, which is different from cross-sectional data by observing subjects at one point of time. Time series analysis is the method to obtain useful statistics and information from time series data, which is similar to regression analysis in the purpose of extracting data by exploratory data analysis, modeling and forecasting. Time series data analysis is applied in many fields, including economics, finance, biology, medicine, etc. This report uses a monthly financial data *sales* of 150 observations taken from Box and Jenkins (1970) in the package *astsa* in R to demonstrate the statistical methods of time series data analysis.

Statistical Methods

Exploratory data analysis

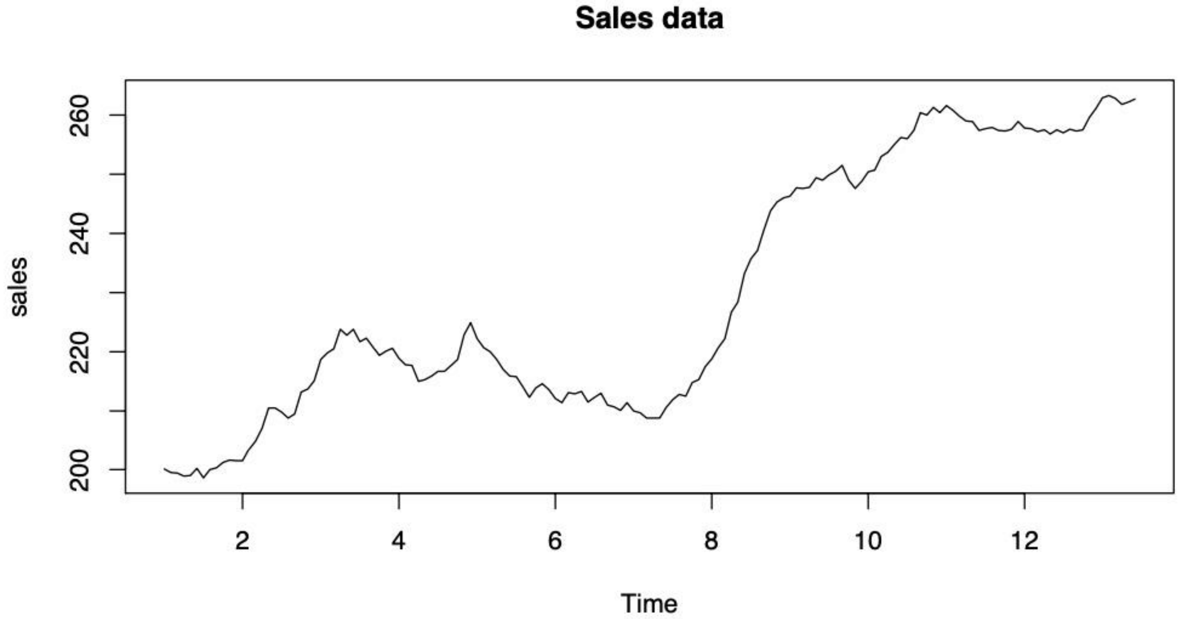


Fig. 1. Plot of time series sales.

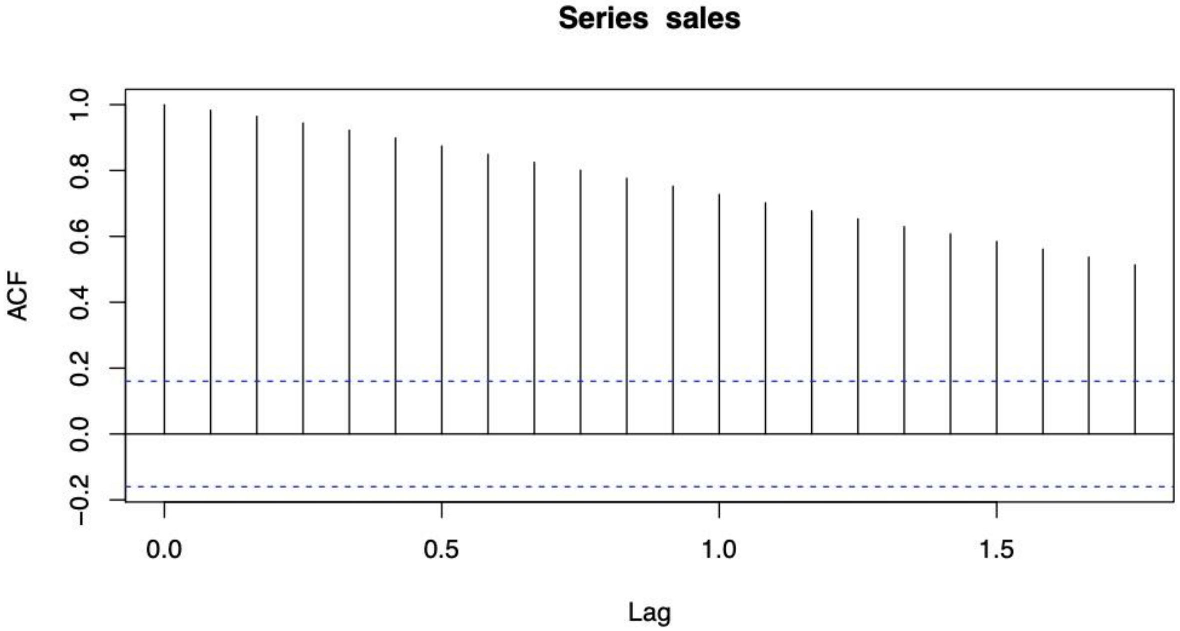


Fig. 2. ACF of time series sales.

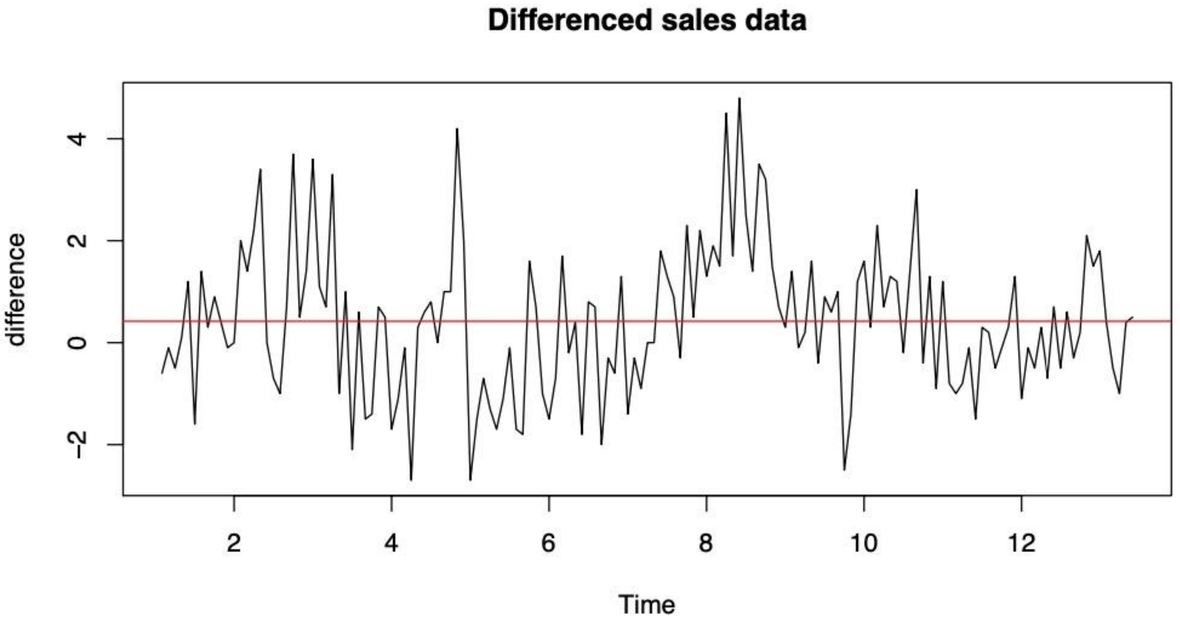


Fig. 3. Plot of one differenced time series sales.

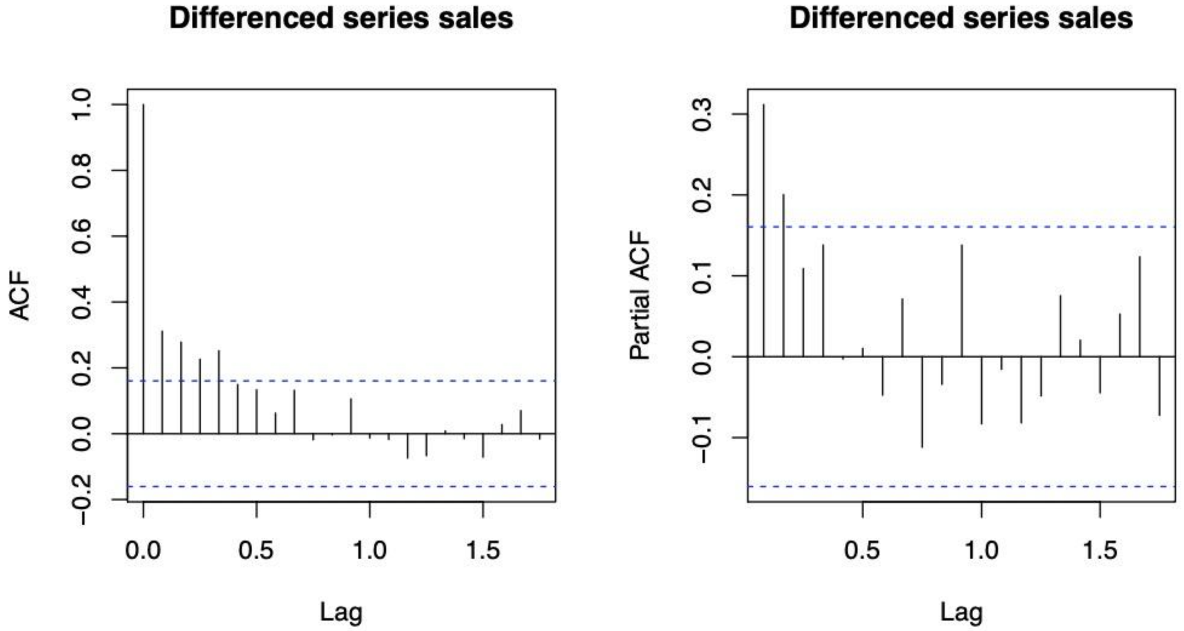


Fig. 4. ACF and PACF of differenced time series sales.

Figure 1. is the plot of time series *sales*, which shows a trend of increasing by time, so the time series is not stationary. Figure 2. is the ACF of time series sales, and shows a trend of slow decay, meaning there is obvious correlation of different points of time even if time lag is large. Thus, one regular difference is suggested. After one difference, the plot of time series *sales* is showed in figure 3., which looks stationary since the time series has constant mean approximately 0.4 and the variance is not changing by time. Figure 4. shows the ACF and PACF of differenced time series *sales*. The ACF shows no trend of slow decay and no seasonal periodic pattern.

Modeling

After one differencing, figure 4. shows that ACF cuts off at lag 4, PACF tails off, suggesting *sales* follows an ARIMA(0,1,4) model. Also, PACF cuts off at lag 2, ACF tails off, suggesting *sales* follows an ARIMA(2,1,0) model.

Results

Model parameter estimates

We define $\hat{x}_t = \nabla\, sales$, and use MLE to fit the MA(4) model. When testing the significance of model coefficients, the p-value of *ma3* is greater than the significance level of 5%, meaning $\widehat{w}_{t-3}$ is not significant, so it is removed from the model. Now we have the following estimated model:

$$\hat{x}_t = 0.413_{(0.182)} + 0.214_{(0.083)}\,\widehat{w}_{t-1} + 0.172_{(0.085)}\,\widehat{w}_{t-2} + 0.154_{(0.073)}\,\widehat{w}_{t-4} + \widehat{w}_t$$

Where $\widehat{\sigma}_w = 1.333$ is on 144 degrees of freedom. The values inside parentheses are the corresponding estimated standard errors.

The estimated AR(2) model for $\hat{x}_t$ is

$$\hat{x}_t = 0.414_{(0.197)}(1 - 0.249 + 1 - 0.199) + 0.249_{(0.080)}\hat{x}_{t-1} + 0.199_{(0.080)}\hat{x}_{t-2} + \widehat{w}_t$$

Where $\widehat{\sigma}_w = 1.339$ is on 146 degrees of freedom, and all of the parameter estimates are significant.

Model diagnostics

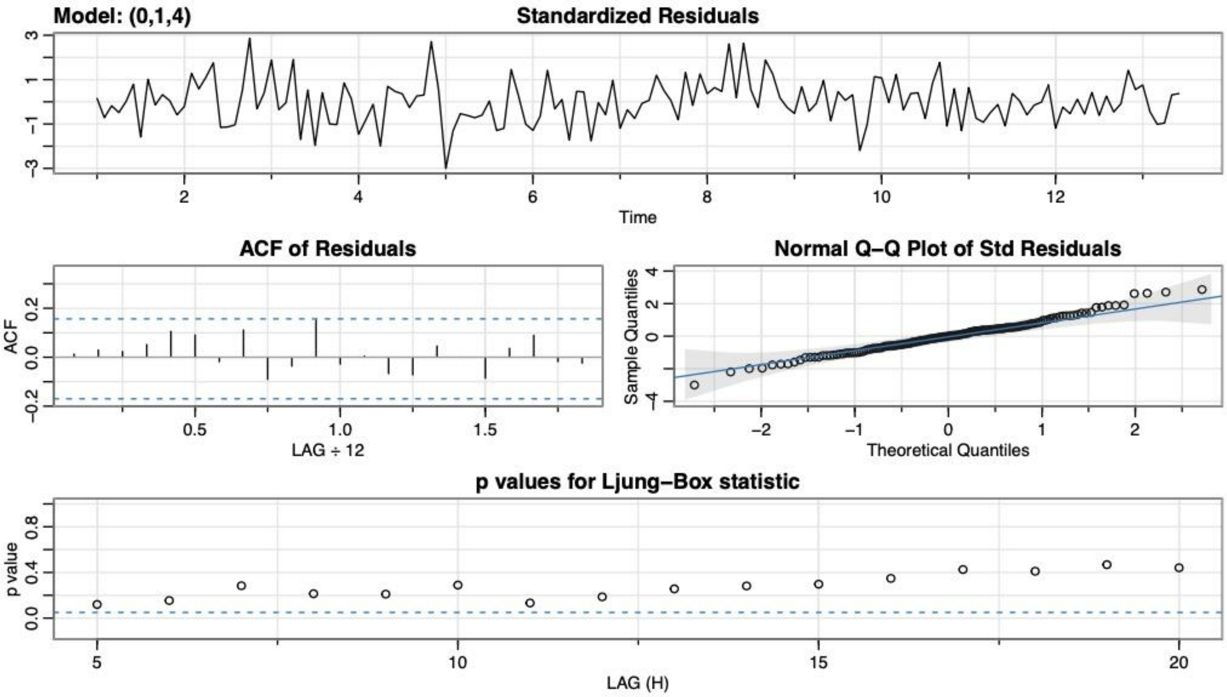


Fig. 5. Plot of ARIMA(0,1,4) model diagnostics.

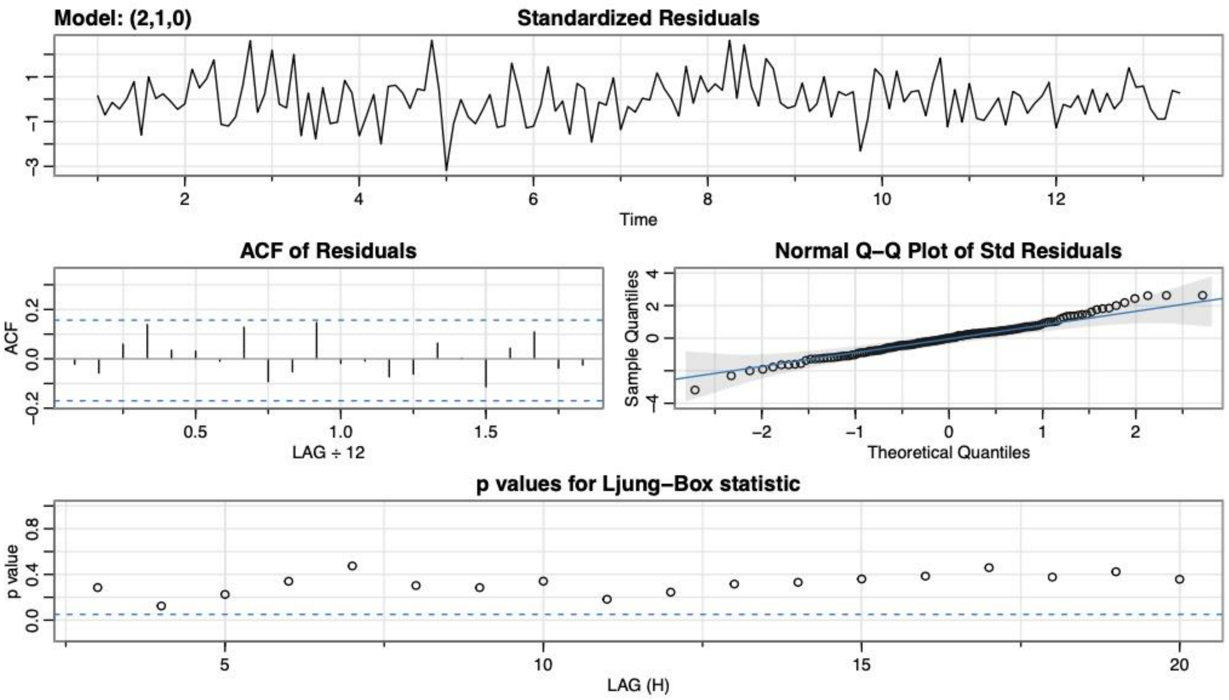
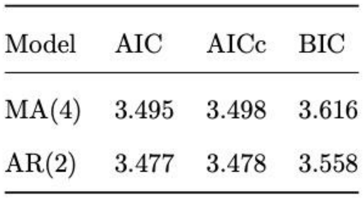


Fig. 6. Plot of ARIMA(2,1,0) model diagnostics.

According to figure 5. & 6., for both models, the standardized residuals plots show no obvious patterns and is constant around 0. The ACF of residuals show no significant spike, so the residuals are close to white noise. The normal Q-Q plots of standardized residuals show that the normality assumption holds. The p-values for Ljung-Box statistic are above the significance level of 5%, so we cannot reject the null hypothesis of Ljung–Box test that ρ(1) = ... = ρ(h). Thus, we say two models pass the model diagnostics and appear to fit well.

Model selection

Table. 1. AIC, AICc, BIC for two fitted models.



Model	AIC	AICc	BIC
MA(4)	3.495	3.498	3.616
AR(2)	3.477	3.478	3.558

Table 1. shows that all of AIC, AICc and BIC of AR(2) model are smaller than MA(4) model, so we choose the ARIMA(2,1,0) model to be the final model.

Forecasting

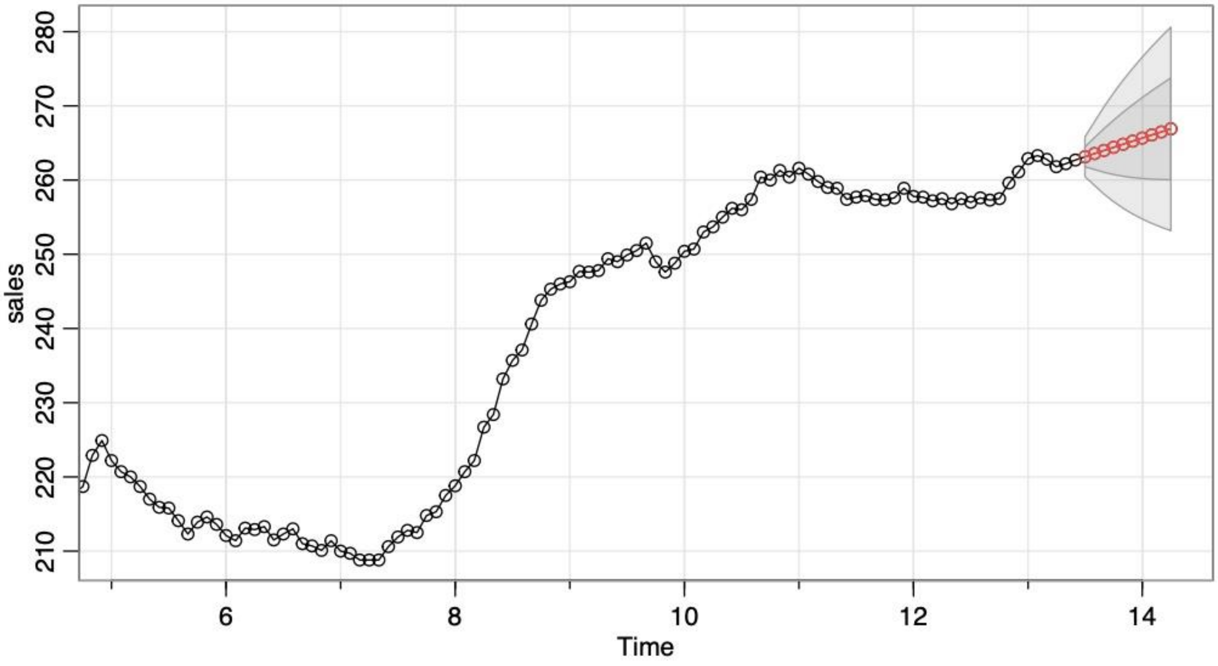
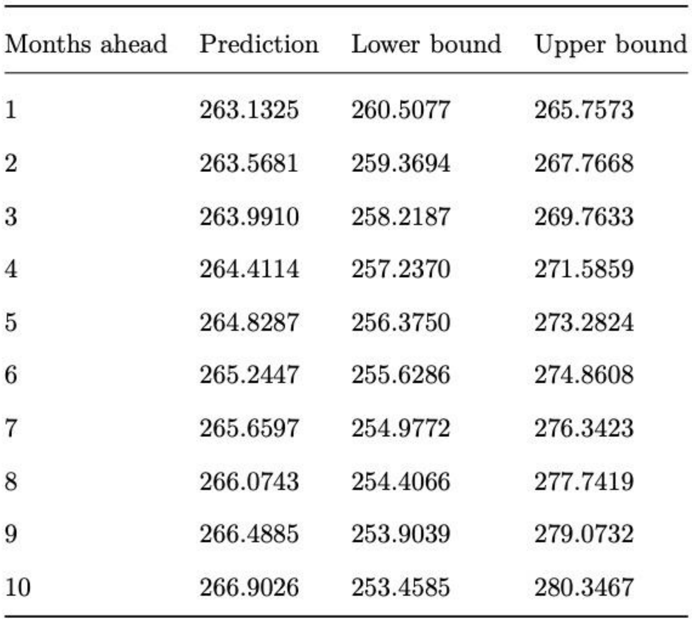


Fig. 7. Plot of forecasting sales into the future ten months ahead.

Table 2. Prediction and 95% prediction intervals for each of the ten forecasts.



Months ahead	Prediction	Lower bound	Upper bound
1	263.1325	260.5077	265.7573
2	263.5681	259.3694	267.7668
3	263.9910	258.2187	269.7633
4	264.4114	257.2370	271.5859
5	264.8287	256.3750	273.2824
6	265.2447	255.6286	274.8608
7	265.6597	254.9772	276.3423
8	266.0743	254.4066	277.7419
9	266.4885	253.9039	279.0732
10	266.9026	253.4585	280.3467

Figure 7. is the plot of forecasting sales into the future ten months ahead, where the red points represent predicted values and light grey area is the 95% prediction intervals. Table 2. shows the corresponding predicted values and 95% prediction intervals for each of the ten months forecasts. The predicted ten months values seem reasonable, especially when we have the 95% PIs that we are 95% sure that the true values will fall into the corresponding intervals showed in Table 2.

Spectral analysis

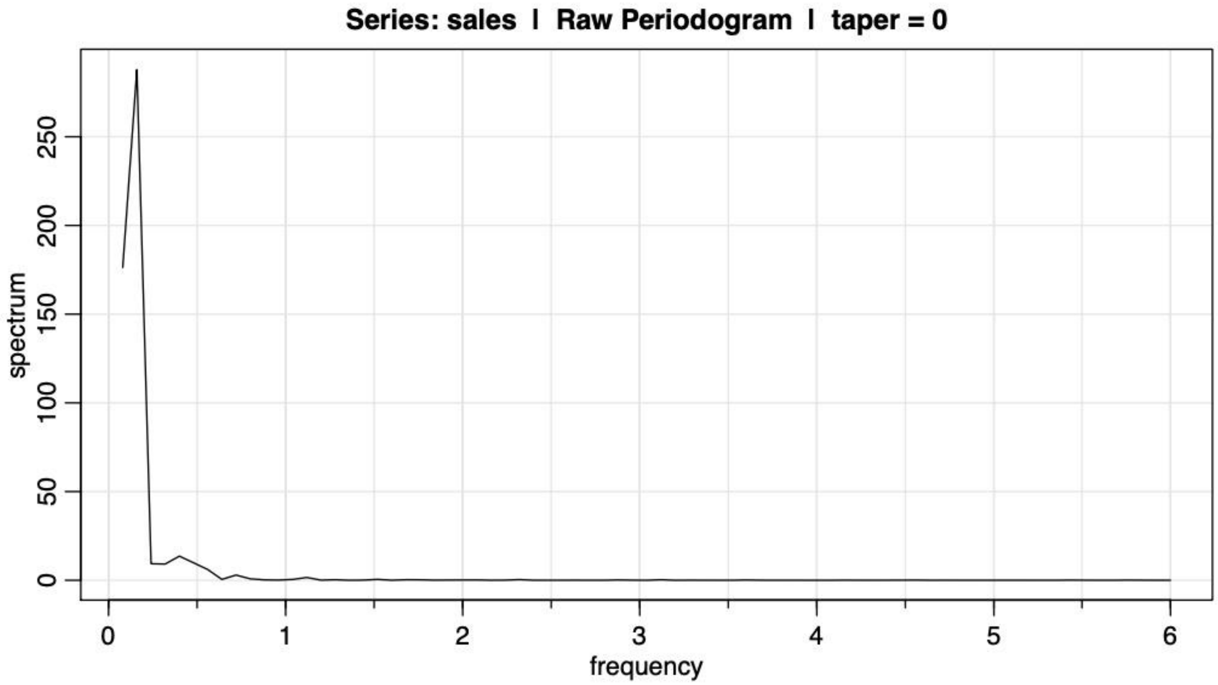
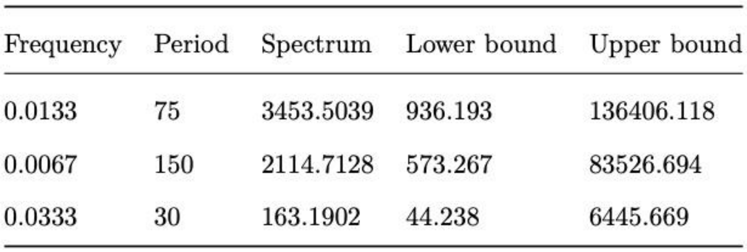


Fig. 8. Plot of the periodogram of sales series.

Table. 3. First three predominant periods and corresponding 95% CI bounds



Frequency	Period	Spectrum	Lower bound	Upper bound
0.0133	75	3453.5039	936.193	136406.118
0.0067	150	2114.7128	573.267	83526.694
0.0333	30	163.1902	44.238	6445.669

We also perform a periodogram analysis. Figure 8. is the plot of the periodogram of the series, showing common peaks at the 6.25 years period (75 months) with $\sigma = 0.0133$. We then identify the first three predominant periods; the results are showed in Table 3. The first dominant frequency has 6.25 cycles occurring years; the second dominant frequency has 12.5 cycles occurring years; the third dominant frequency has 2.5 cycles occurring years. We are 95% sure that the dominant frequencies will fall into the corresponding intervals showed in Table 3. The 95% CIs are significantly wide, so we cannot establish significance of the peak.

Discussion

We demonstrate the statistical methods of time series analysis through *sales* series by exploratory data analysis at first. Then we suggest statistical models based on exploratory analysis, and select best fitted model which has better performance by model diagnostics to predict future values; finally, we perform spectral analysis to study the periodic signal of the series. There are some limitations in this study; it is relatively subjective to choose the parameters for the ARIMA model. The ACF and PACF of differenced *sales* series showed in figure 4. may also suggest ACF cuts off at lag 5, PACF tails off, meaning ARIMA(0,1,5) model would also be considered. Besides, the simpler model ARIMA(1,1,1) or ARIMA(1,1,0) or ARIMA(0,1,1) are not being considered, which may sometimes have better model performance. Thus, in the future studies, more statistical models can be suggested to select best fitted model for forecasting times series.

Reference:

David Stoffer (2021). *astsa: Applied Statistical Time Series Analysis*. R package version 1.14. https://CRAN.R-project.org/package=astsa