Seasonal time series analysis: Monthly number of sales

Author: Ziqi Fan

Date: 2023-04-06

Introduction

The Time Series Data Library (TSDL) was developed by Rob Hyndman, a statistics professor at Monash University in Australia, to provide a collection of time series datasets for analysis. Among the many subjects covered in TSDL, I will be using sales data to build the most accurate model possible for predicting future sales. The dataset contains monthly mean relative sales figures from the years 1965 to 1975.

Methodology

Exploratory Analysis

The dataset contains information about the monthly sales figures, which were recorded from January 1965 to December 1975. The are 132 records in total, with a median of 44.00 and a mean of 45.36. The smallest value is 23.00 and the maximum value is 72.00, and there are no missing values. To visualize any trends, seasonality, or other patterns in the data over time, I plotted the time series. However, as the original data contains many fluctuations, it can be difficult to interpret. To address this, I plotted the moving average to smooth out the noise and highlight any underlying trends or patterns. Furthermore, I used the decomposition of an additive time series to break down the time series into its four component parts: the original data, trend, seasonality, and random noise. This approach provides a better understanding of the underlying patterns and trends within the time series and can help identify any factors contributing to those patterns."

Modeling

To identify the best model for analyzing the time series data, I considered four models: ARIMA, seasonal ARIMA, seasonal exponential smoothing (SES), and a fourth model, which is not mentioned. As this is a seasonal time series dataset, I selected models that can account for seasonal fluctuations. Before modeling, I split the original data into approximately 70% training and 30% testing sets to obtain more accurate results. The training set consisted of data from the beginning of the dataset to the end of 1972, while the testing set comprised data from the beginning of 1973 to the end of the dataset.

The first model considered for analyzing the time series data is ARIMA, which is particularly effective in modeling data with regular patterns of seasonal fluctuations, such as the monthly sales numbers. Since the sales cycle has a regular pattern of one year, an ARIMA model that incorporates a seasonal component can be used to capture this pattern. However, before applying the ARIMA models, it's crucial to ensure that the time series is stationary, meaning that its statistical properties like mean and variance don't change over time. To identify the parameters p, d, and q in the ARIMA model, the ACF and PACF are used, and a guideline for selecting the model is shown in **Figure 1**. The second model considered is the Seasonal ARIMA (SARIMA) model, which is an extension of the standard ARIMA models and includes additional terms to account for the seasonal component of the data. The third model to be considered is the seasonal exponential smoothing (SES) model, which is an extension of the single exponential smoothing model. The SES model takes into account seasonal factors and adjusts the weights for each

season of the year. Before applying the SES model, it's necessary to estimate the values of the smoothing parameters and the seasonal cycle that maximize the likelihood of the observed data. This can be achieved by using the previously automatically selected alpha in the SARIMA model, which is a method for finding the parameter values that make the observed data most probable given the model. To compare the performance of the different forecasting methods, accuracy measures are calculated using the test data. It's important to note that the choice of the best model depends on the specific forecasting problem and the measure of forecast accuracy that is most important for the decision-maker.

Model	ACF	PACF
ARIMA(p, d, 0)	Trails off to zero	Zero after lag p
ARIMA(0, d, q)	Zero after lag q	Trails off to zero
ARIMA(p, d, q)	Trails off to zero	Trails off to zero

Figure 1. Guidelines for selecting an ARIMA model

Analysis

Exploratory Analysis

The time series data of the number of sales does not exhibit any noticeable long-term increasing or decreasing trend. However, it displays a non-distinctive pattern of two peaks that occur approximately every month in March and August, indicating a seasonal pattern to the sales cycle (as shown in **Figure 2**). Although the intensity of these peaks may vary, the regularity of the cycle implies that there is a reason behind the sales fluctuations. Studying the sales cycle and its underlying causes is a crucial research area in business and has practical implications for companies.

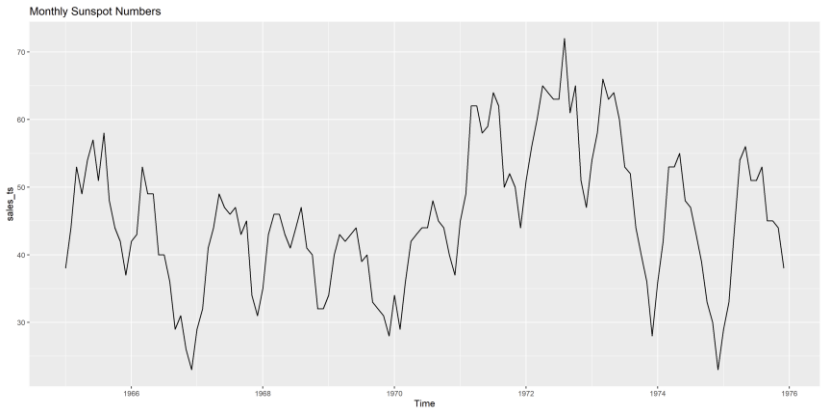

Figure 2. Monthly sales Numbers

Based on the decomposition of additive time series, the sales dataset has been separated into four components - data, trend, seasonal, and remainder - to analyze its pattern. In **Figure 3**, there are distinct recurring patterns that indicate the presence of seasonality. Specifically, there are two

peaks in sales every March and August, and the sales reach their lowest point at the end of every month.

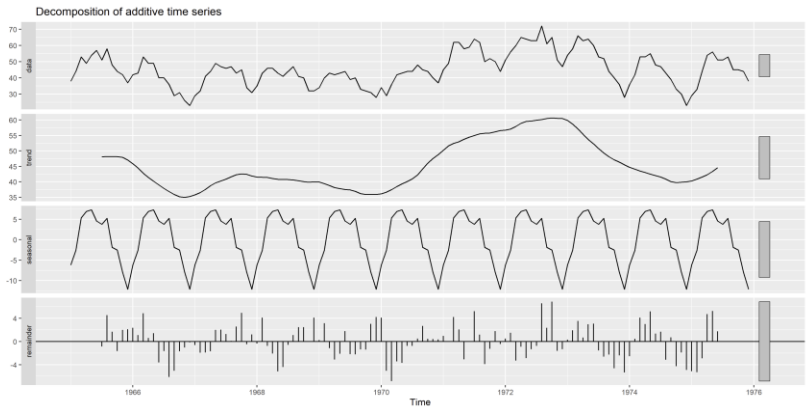

Figure 3. Decomposition of additive time series

Main data analysis

ARIMA

The initial step was to fit an ARIMA model to the training set, and the KPSS Test was used to verify the stationarity of the data. However, the obtained p-value (0.07037) was greater than the significance level, indicating that the data was non-stationary. In order to focus on the seasonal patterns, the seasonal component of the sales data was removed, and a new KPSS Test was conducted. This time, the p-value (0.02495) was less than 0.05, indicating that the data was stationary. Next, ACF and PACF plots were plotted, and based on the criteria discussed previously (in Figure 1), an ARIMA (1, 1, 0) model was selected since the ACF trails off to zero and the PACF is zero after lag p.

After fitting the model, a summary was generated, which showed that the sigma square was 10.83, and the log-likelihood was -247.97, resulting in an AIC of 499.94. The model utilized only the AR1 coefficient. Additionally, error measures were calculated for the training set, which showed that both the ME and MPE values were close to zero, indicating that the model's errors were relatively small and unbiased. The RMSE and MAE values indicated that the model had an average error of about 3.27 and 2.64, respectively, on the training set. The MAPE value was about 5.99%, indicating that the model's average absolute percentage error was about 5.99%. The MASE value was about 0.98, which indicated that the model's accuracy was relatively good compared to a naive model that predicts the next value based on the most recent observation. Finally, the ACF1 value was about -0.014, indicating that the model's residuals had some autocorrelation at lag 1, but this was not a major concern since it was close to zero.

SARIMA

In this case, I choose to let the R system choose the best model for SARIMA, and it turns off is ARIMA (1, 0, 2) (1, 0, 0)[12]. The model has two coefficients non-seasonal MA term (ma1) and one seasonal MA term (sma1) with corresponding coefficients -0.2528 and -0.1960, respectively.

The standard errors of the estimated coefficients are also provided. The estimated variance of the error term (sigma^2) is 10.69, and the log-likelihood of the model is -246.61. The AIC, AICc, and BIC values are 499.21, 499.48, and 506.87, respectively. The training set error measures indicate that the model has a mean error (ME) of 0.245236, a root mean squared error (RMSE) of 3.218307, a mean absolute error (MAE) of 2.65721, a mean percentage error (MPE) of 0.09754921, and a mean absolute percentage error (MAPE) of 6.025415. The MASE (mean absolute scaled error) is 0.3504013, and the autocorrelation of the residuals at lag 1 (ACF1) is -0.01460636.

Seasonal exponential model

The output indicates that the estimated smoothing parameter alpha is 0.7596, which determines the weight given to the most recent observation in the smoothing process. The initial level l is estimated to be 44.2174, which is the initial smoothed value of the series. The estimated standard deviation of the error term is given by sigma, which is 3.2689. The output also provides several error measures for evaluating the model's performance on the training set. The mean error (ME) is 0.2100648, indicating that, on average, the predictions are slightly higher than the actual values. The root mean squared error (RMSE) is 3.268931, indicating that the model's predictions have an average error of around 3.27 units. The mean absolute error (MAE) is 2.653659, indicating that, on average, the predictions are off by around 2.65 units. The mean percentage error (MPE) is 0.06047718, indicating that, on average, the predictions are slightly higher than the actual values. The mean absolute percentage error (MAPE) is 6.0404%, indicating that, on average, the predictions are off by around 6.04%. The mean absolute scaled error (MASE) is 0.3499331, indicating that the model is performing reasonably well compared to a naive forecast. Finally, the autocorrelation of the residuals at lag 1 (ACF1) is -0.003876693, suggesting that there is no significant correlation between residuals at adjacent time points.

Discussion

The report discussed five models: ARIMA, SARIMA, Seasonal Exponential Model, and Holt-Winters Exponential Smoothing. **Figure 4** compares the accuracy of these three models based on error measures. We can see that the SARIMA model has the lowest values for ME, RMSE, MAE, and MASE for the test sets, indicating that it has the lowest overall forecast error among the three models. In terms of forecast accuracy, we can look at the MAPE (Mean Absolute Percentage Error) values, which show that the SARIMA model has the lowest MAPE value for the test set, indicating that it has the lowest percentage error in forecasting the test data. Therefore, I choose the SARIMA model as the best-performing model among all three. **Figure 5** shows the forecasted sales numbers for the original data for the next 12 months.

Model	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
ARIMA	-12.914	14.689	13.145	-30.887	31.269	1.733	0.845	5.168
SARIMA	-12.420	14.157	12.609	-29.734	30.046	1.663	0.848	4.996
Single exponential model	-13.341	15.065	13.500	-31.832	32.096	1.780	0.845	5.293

Figure 4. The accuracy of the three models on test set

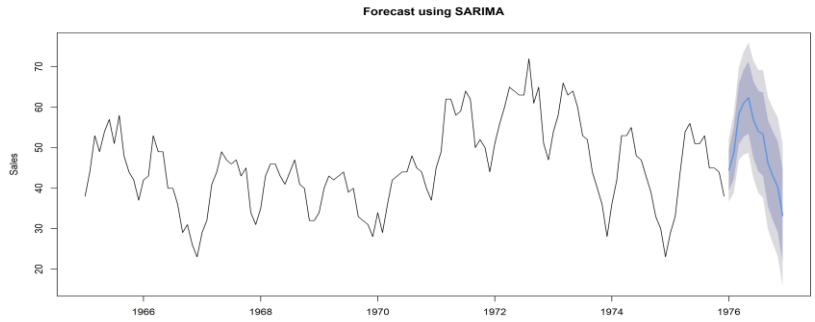

Figure 5. The forecast sales number in the next 12 months.

Conclusion

There are several ways to choose models, such as AIC, ROC curves, or accuracy measures. The choice of the best model for forecasting future data depends on the method used to evaluate the models. In this report, I have used the accuracy of the test data as the criterion for selecting the best model. Based on this criterion, the seasonal ARIMA model is the best fit for the sales dataset. The forecast indicates a relatively high sales number in the coming years, with a sharp increase followed by a steep decline. Therefore, it is important to ensure sufficient inventory for sales before March 1976.