Time Series Forecasting with SARIMA: Detailed Analysis

Jesús Veiga Morandeira

2024-12-27

Table of contents

1 Introduction to Time Series Analysis

Time series analysis is a statistical technique that deals with the analysis of data points collected or recorded at specific time intervals. Unlike other forms of data analysis, time series data exhibits a temporal ordering. This temporal ordering is important, as time series analysis aims to understand the underlying structure and function of the data to make forecasts or inform decisions.

2 Mathematical Foundations of Time Series Models

A time series (y_t) is a sequence of observations indexed by time t , where (t = 1, 2, ..., T). Time series models aim to capture the relationship between observations using mathematical equations. Some of the key models include:

2.1 1.1.1 Autoregressive (AR) Models

An AR(p) model expresses the current value (y_t) as a linear combination of its (p) previous values:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$$

where:

- c is a constant,
- (ϕ_1, ϕ_2, ..., ϕ_p) are autoregressive coefficients,
- ε_t is white noise.

2.2 1.1.2 Moving Average (MA) Models

An MA(q) model represents (y_t) as a function of past error terms:

$$y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

where μ is the mean, and (θ_1, θ_2, ..., θ_q) are the moving average coefficients.

2.3 1.1.3 ARIMA Models

The (ARIMA(p, d, q)) model generalizes AR and MA models to handle non-stationary data by introducing differencing (d):

$$\nabla^d y_t = (1 - B)^d y_t,$$

where (B) is the backshift operator.

2.4 1.1.4 Seasonal ARIMA (SARIMA) Models

SARIMA extends ARIMA to incorporate seasonality. The SARIMA model is denoted as SARIMA$(p, d, q)(P, D, Q)_s$, where:

- (P, D, Q) are the seasonal orders,
- s is the seasonal period.

The general equation is:

$$\Phi(B^s)\phi(B)(1 - B)^d(1 - B^s)^D y_t = \Theta(B^s)\theta(B)\varepsilon_t.$$

These models are essential tools in fields like economics, meteorology, and engineering to understand and predict time-dependent phenomena.

3 Data Description

The dataset analyzed in this document contains monthly air passenger counts from January 1949 to December 1960. It is a well-known benchmark dataset for time series analysis.

3.1 Dataset Summary

- **Time Period**: January 1949 - December 1960.
- **Frequency**: Monthly.
- **Variables**:
  - `Month`: Time period in `YYYY-MM` format.
  - `#Passengers`: Total number of passengers for the respective month.

3.1.1 Statistical Overview of Passengers

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  104.0   180.0   265.5   280.3   360.5   622.0 
```

Key statistics:

- **Mean**: 280.30 passengers.
- **Standard Deviation**: 119.97 passengers.
- **Minimum**: 104 passengers.
- **Maximum**: 622 passengers.

The dataset shows an increasing trend and clear seasonality, making it ideal for SARIMA modeling.

4 Methodology

4.1 Splitting the Data

The dataset was split into training (80%) and testing (20%) sets to evaluate the model's performance on unseen data.

4.2 SARIMA Model Specification

A SARIMA$(1,1,0)(1,1,0)_{12}$ model was chosen based on the dataset's characteristics:

- **Non-stationarity**: Handled by first-order differencing (d=1).
- **Seasonality**: Captured by seasonal differencing (D=1) and seasonal lags (P=1, Q=0).

4.2.1 Model Fitting and Forecasting

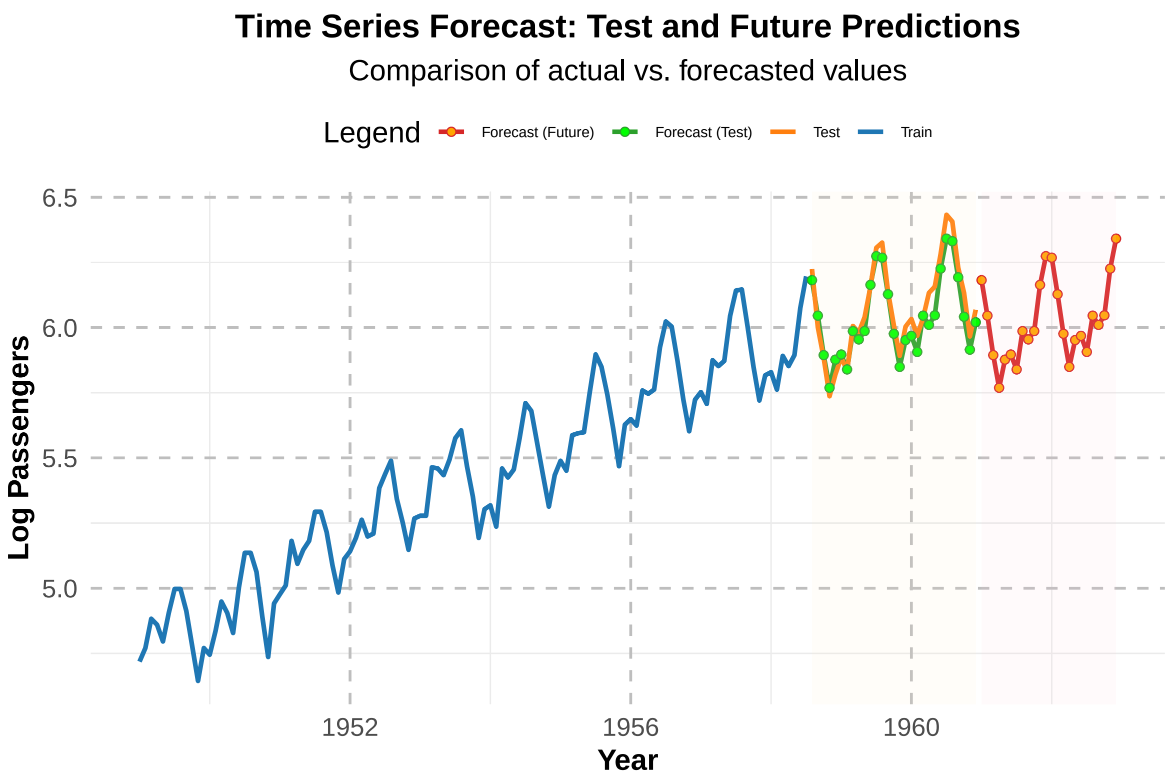
We used the `sarima.for` function from the `astsa` package to forecast both the test set and future values.

5 Results and Visualization

The forecasts were visualized alongside the actual data for better interpretability. The graph highlights:

- Training data in blue.
- Testing data in orange.
- Forecasted values for the test set in green.
- Forecasted future values in red.

5.0.1 Visualization Code



5.0.2 Final Observations

- The model accurately captures both the trend and seasonality in the training data.
- Predictions for the test set align closely with actual values, indicating good model performance.

- Future forecasts demonstrate a continuation of the upward trend and seasonal pattern observed in the data.