Rain in Australia

Forecasting the amount of rainfall, in mm, using the Box-Jenkins Methodology

Carolina Almeida 20221855, João Capitão 20221863, Teresa Simão 20221873

Introduction

From tropical downpours to arid landscapes, understanding and predicting rain patterns is essential for managing resources and ensuring resilience in the face of changing climates. In Australia, the significance of this task transcends various sectors, spanning agriculture, water resource management by optimizing water usage, and public safety by helping mitigate the impacts of floods and droughts. This study contributes to the ongoing efforts in climate science to improve predictive capabilities and support informed decision-making in weather-dependent sectors.

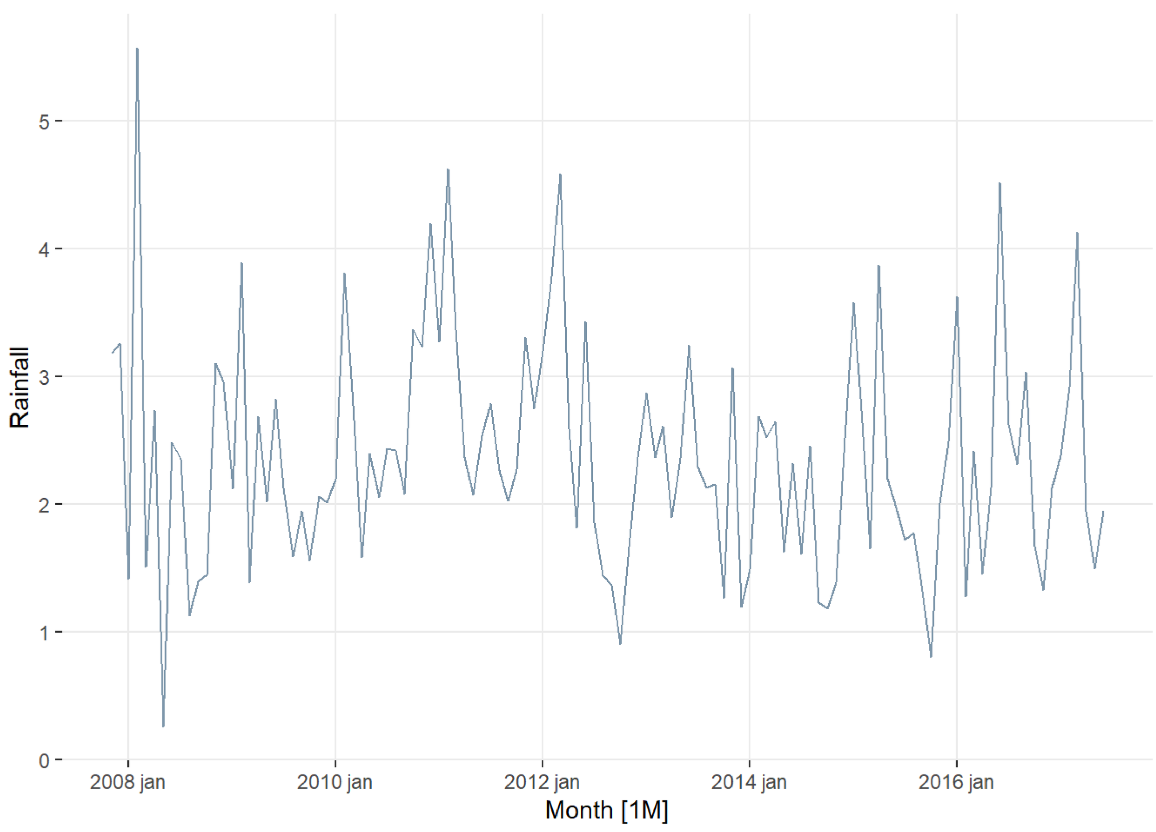

Figure 1: Daily amount of rainfall (mm)

Methodology

Initially, the gaps and missing values were addressed. The latter was managed through a systematic approach: replacing missing values with the mean (since it maintained the dataset's central tendency and preserved the overall data structure). Following this, the dataset was divided into two subsets: a validation set encompassing years from 2008 to 2014, and a test set containing data from 2015 to 2017. Our goal was to achieve a balanced split of 75% for training and 25% for testing, a target we successfully met.To stabilize the variance of our training time series, we applied a **Box-Cox transformation**. Recognizing that the data still lacked stationarity, we then implemented a seasonal differencing technique. The resulting series was confirmed to be stationary through a **unit root test**.

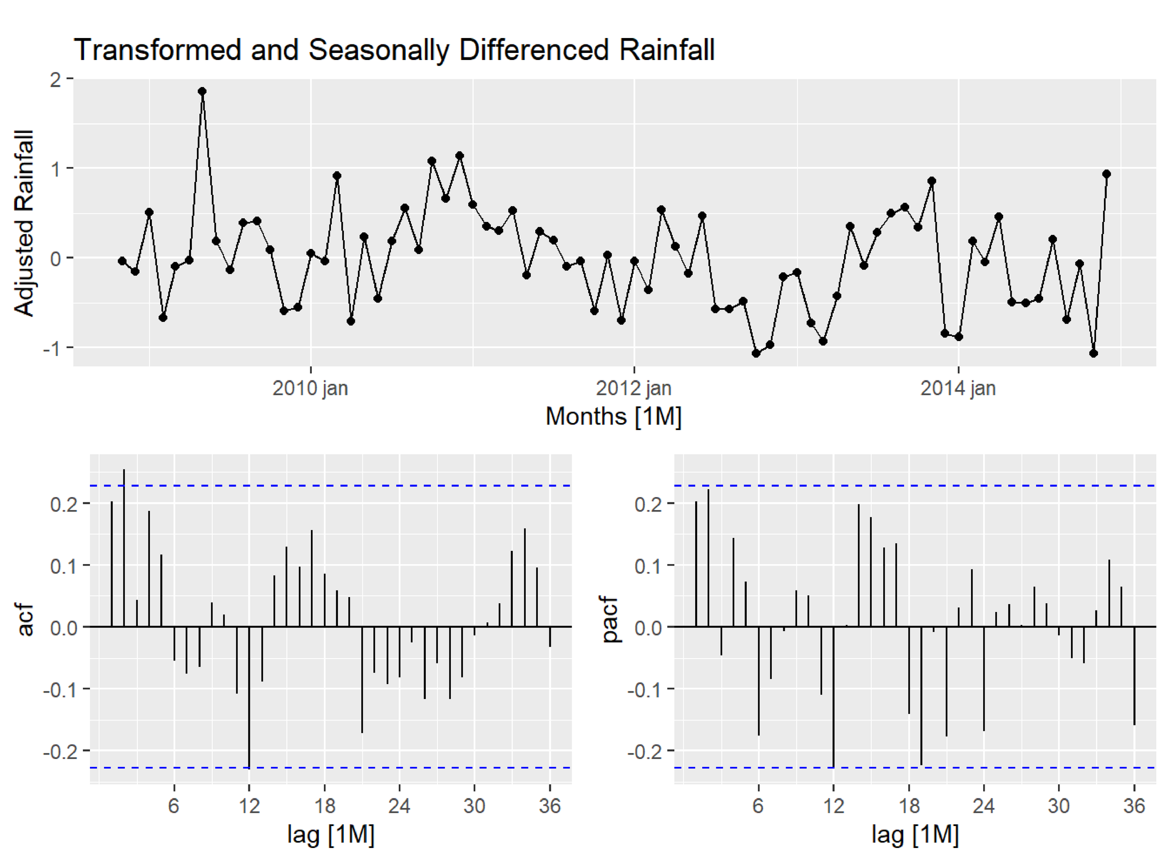

Figure 2: Daily amount of rainfall (mm) after differentiation; ACF Plot; PACF Plot

Based on the autocorrelation and partial autocorrelation plots obtained, several ARIMA models were selected and evaluated, using the following criteria:

- d = 0 since differencing was not applied to achieve stationarity
- D = 1 since seasonal differencing was applied
- p is typically set to 2 due to a spike at lag 2 in the PACF plot
- q is sometimes set to 2 due to a spike at lag 2 in the ACF plot
- P is generally set to 0 due to the absence of noticeable spikes at lags 12, 24, and 36 in the PACF plot
- Q is usually set to 1 due to a spike at lag 12 in the ACF.

Additionally, the **Ljung-Box test** was applied to these models, and none showed statistical evidence of autocorrelation in the residuals.

Final Results

After completing the evaluation and analyzing the errors, illustrated bellow, it was concluded that SARIMA(2,0,0)(1,1,1) and SARIMA(2,0,0)(0,1,2) were the best models.

.model	.type	ME	RMSE	MAE	MPE	MAPE
auto_ARIMA	Test	-0.0199294	0.9320479	0.7464372	-17.630019	37.78818
sarima002010	Test	0.0769689	1.0990466	0.8505626	-10.426525	39.35719
sarima002011	Test	-0.1222546	0.9621217	0.7733463	-20.663831	39.25458
sarima200010	Test	0.0983372	1.0954496	0.8409188	-9.459692	38.70344
sarima200011	Test	-0.0694643	0.9459729	0.7565522	-18.366968	38.04851
sarima200012	Test	-0.0670657	0.9439404	0.7564593	-18.195045	37.99511
sarima200111	Test	-0.0649901	0.9427952	0.7563558	-18.062238	37.95325
sarima202011	Test	-0.0734147	0.9479260	0.7572750	-18.620852	38.14798

It can be inferred that both of these models effectively captured the historical patterns of rainfall in Australia over recent years.

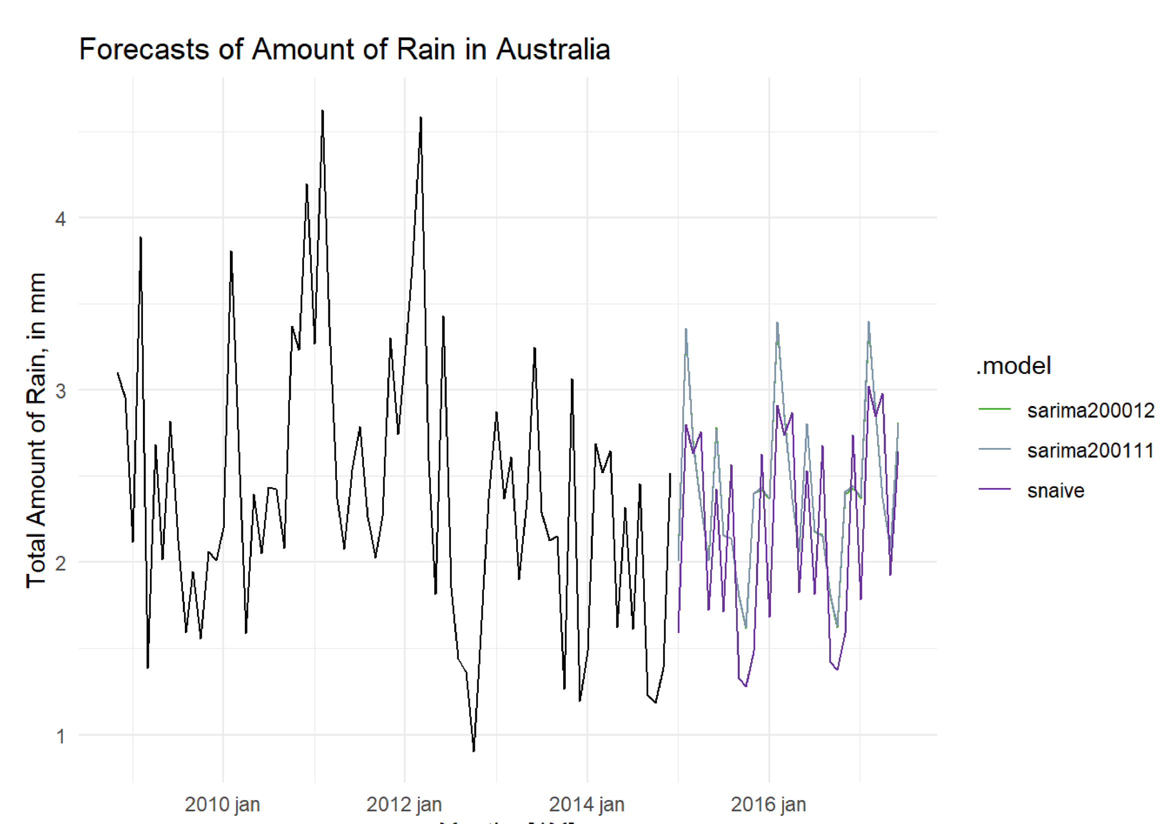

Figure 3: Forecast of amount of rain in Australia - Seasonal Naive; SARIMA(2,0,0)(1,1,1); sarima(2,0,0)(0,1,2)

Conclusion

Among the models considered, SARIMA(2,0,0)(0,1,2) emerged with the lowest RMSE (0.944) and MAE (0.756), indicating superior performance in terms of these error metrics. Additionally, the ME value being very close to zero (-0.067) suggests minimal bias in the predictions. On the other hand, SARIMA(2,0,0)(1,1,1), also demonstrated commendable performance with low RMSE (0.943) and MAE (0.756) values. Also, similar to SARIMA(2,0,0)(0,1,2), its ME and ACF1 values are close to zero, indicating satisfactory model performance.

Thus, the selected model is **SARIMA(2,0,0)(1,1,1)**.

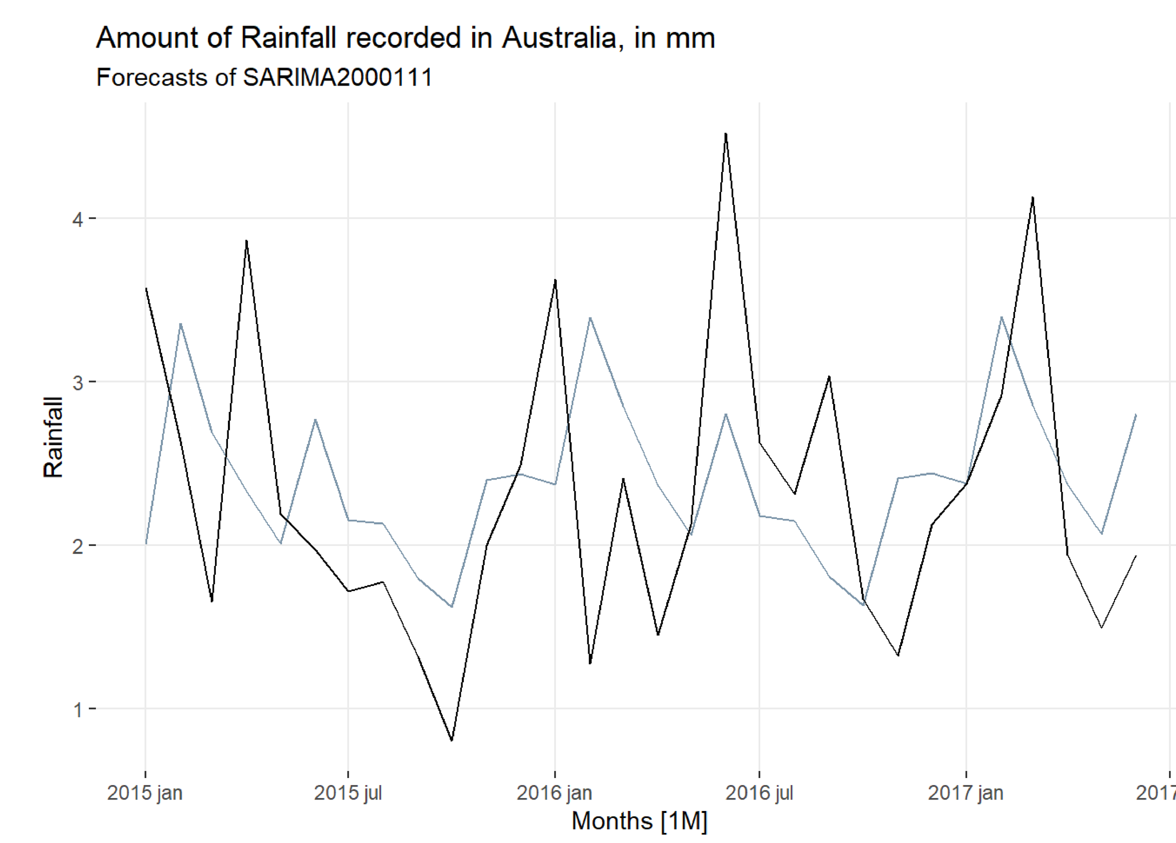

Figure 4: Forecast of amount of rain in Australia

It is important to note that we analyzed the residuals of our model using a plot, and confirmed that they exhibit characteristics of white noise and follow a normal distribution.

To conclude, despite the challenges posed by the dynamic nature of climate, We were able to successful attain a robust and consistent model to forecast rainfall. Our findings underscore the importance of employing sophisticated time series analysis techniques in tackling the complexities of climate prediction. By leveraging SARIMA models, we equip stakeholders with valuable insights to inform decision-making and resource management strategies in the face of Australia's ever-changing weather patterns.

References

[1] Time Series Basics and Getting Started with R

[2] Forecaster's Toolbox

[3] Introduction to Time Series. ARMA models

[4] Box - Jenkins methodology. Forecasting with ARIMA models. Introduction to Seasonal ARIMA models

Data acquired from: https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package