Analysis of Daily Temperatures in Melbourne

1 Introduction

This data analysis report will focus on the structure of the time series data set consisting of daily maximum temperatures (in^oC) in Melbourne and provide the time series model and forecast for time points at the end of the series.

1.1 Background

The information used for this analysis is assigned by **Professor Kostas** [4] and it is kindly provided by the *Time Series Data Library* and the data provider *DataMarket* (DataMarket.com).

The data in this project highlights the maximum daily temperature in Melbourne, Australia, over ten years (1981-1990). There are 3650 observations and the units are in degrees Celsius.

The data file is made up of two columns:

1. Date
2. Daily.maximum.temperatures.in.Melbourne..Australia...1981.1990

Table 1: *Temperature Summary Statistics*

	Date	Temperature
Minimum Temperature	1984-07-02	7^oC
Maximum Temperature	1982-01-24	43.3^oC

Table 1 is showing the maximum and the minimum temperature recorded over 10 years in Melbourne. In January'1982 43.3^oC temperature was recorded which was the minimum temperature from 1981-to 1990 and in July'1984 7.0^oC was recorded as the minimum temperature. Further Analysis defining the structure of the series is done in the EDA section.

1.2 Overview

To determine the structure of the data in section 2, the series is subjected to exploratory data analysis. In Section 3, the Seasonal Autoregressive Integrated Moving Average (SARIMA) model is used to build a model to predict future temperature values. Future maximum temperature values in Melbourne are anticipated in Section 4 using the model selected for SARIMA. Finally, some tests are performed to check that the model is adequate.

2 Exploratory Data Analysis

The provided temperatures were plotted to determine the data's structure. Figure 1 depicts the cyclic pattern in the series but without any presence of a trend over the years.

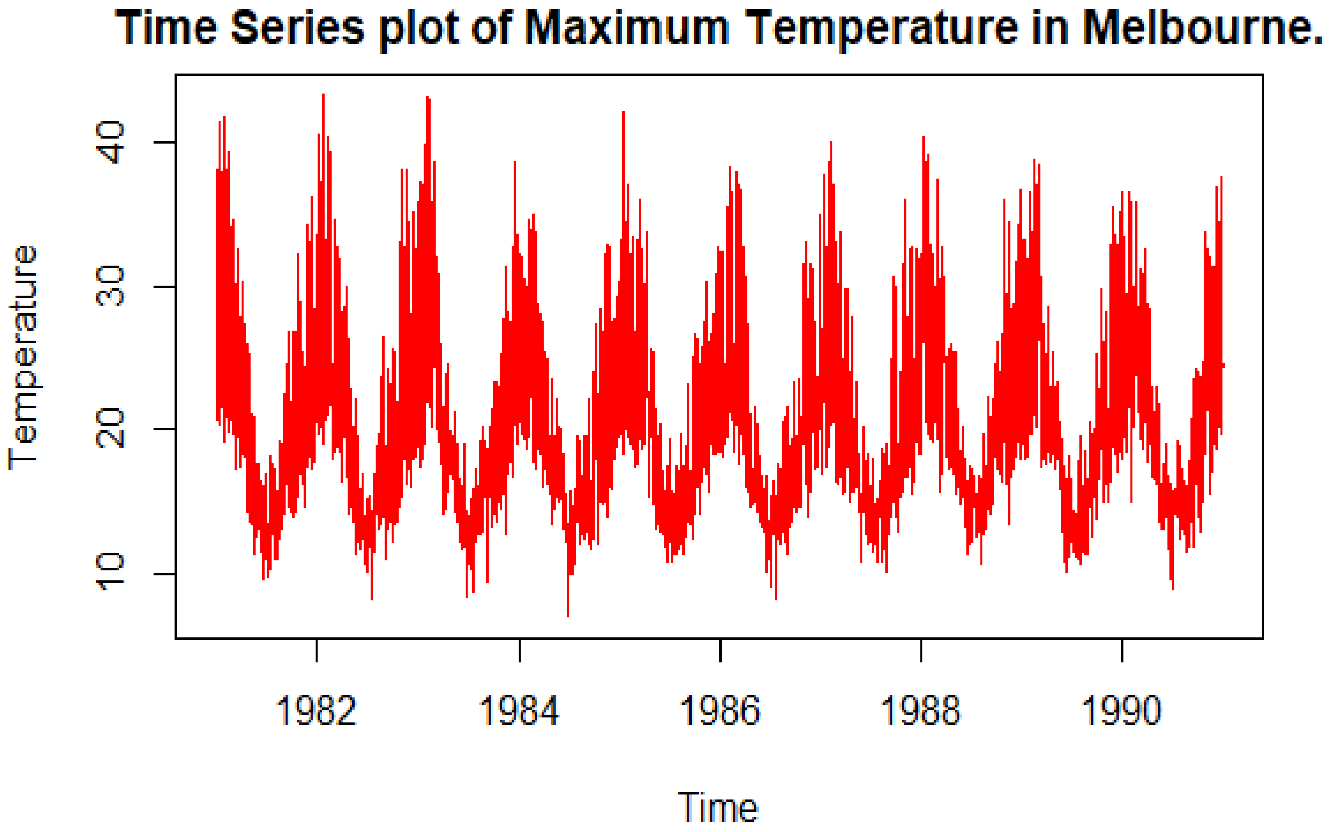

Figure 1: *Time Plot of the Daily Maximum Temperature (°C) in Melbourne showing the consistent seasonal variation over the period.*

Data was decomposed to verify whether there is a trend and confirm the presence of seasonal variation. Additive model decomposition in R was performed [1] on the data since there is no evident proof of increase in seasonal amplitude. Figure 2 depicts that there is no trend during those 10 years but seasonal variation is significant.

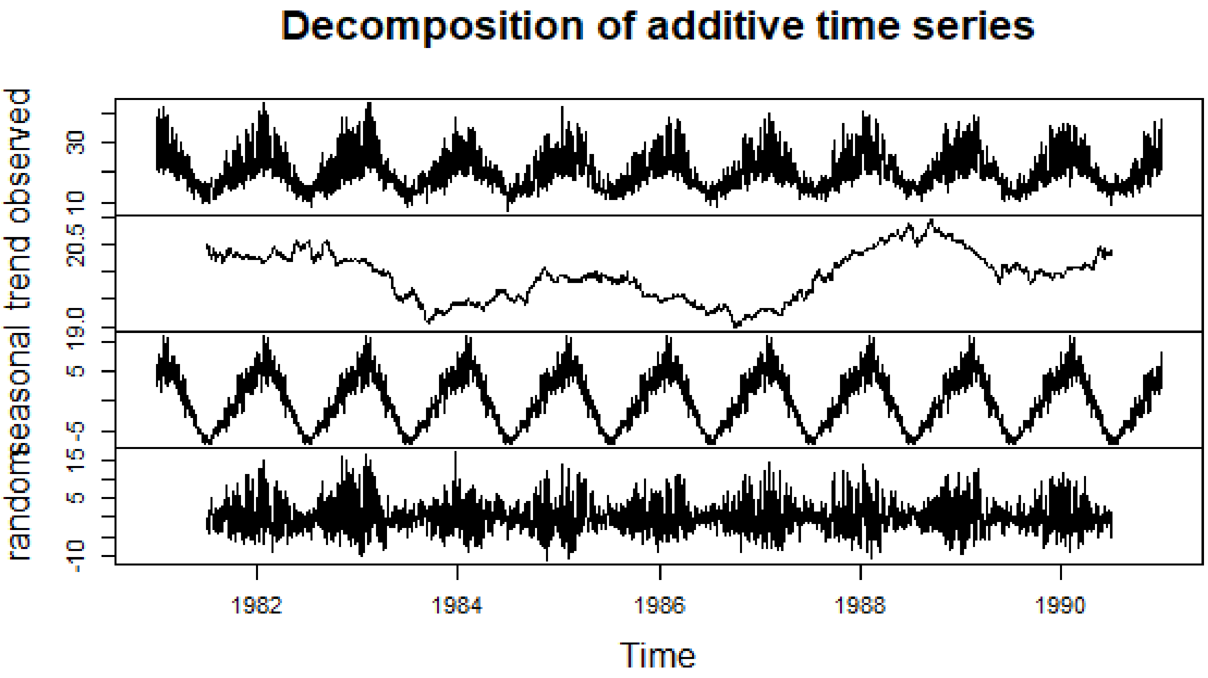

Figure 2: *Graphical Illustration of the Decomposition of the Series. No evidence of trend but seasonal variation can be seen in the series.*

To detect a pattern in the maximum temperatures, sample autocorrelation plot can be used. It will also be used to check whether the series is stationary.

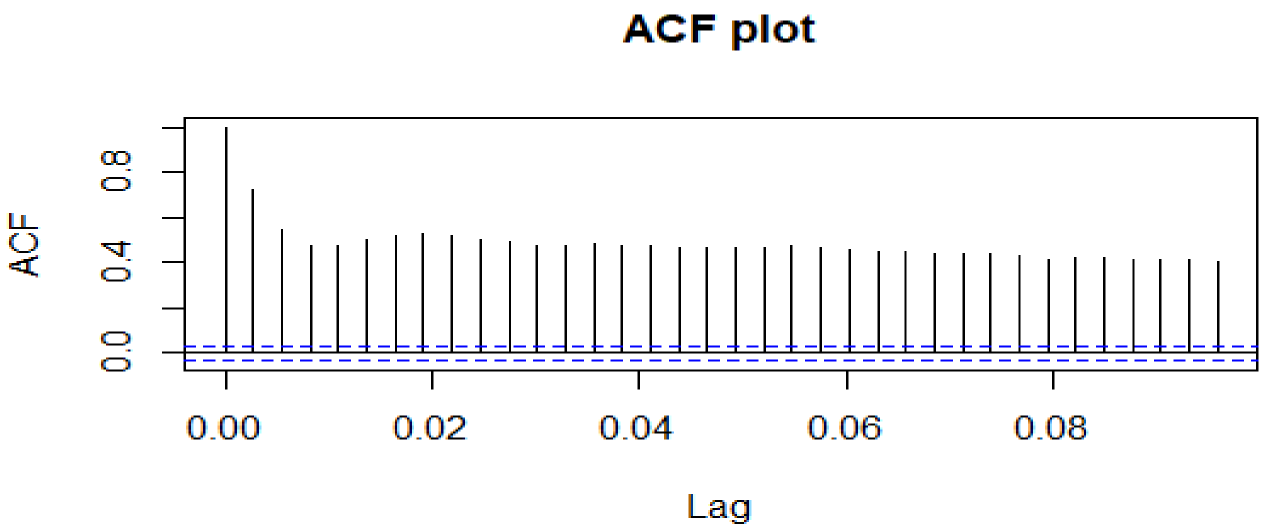

Figure 3: *ACF plot. Failure to decay to zero suggests non-stationary.*

For lag 0, sample autocorrelation is always 1 as seen the Figure 3. The sample correlation between the maximum temperatures in Melbourne are beyond the 95% confidence interval indicating the the daily maximum temperature series is not stationary. Because the series is not stationary, it must be differentiated to make it stationary. Although it is possible to differentiate the series at lag = 365 (since the temperature is measured daily) in RStudio, the main goal here is to fit the data and predict future temperature, and based on EDA, Seasonal Autoregressive Integrated Moving Average (SARIMA) appears to be a good fit (which will be discussed later), and it is not possible to difference the series at lag = 365 to fit the SARIMA model. As a result, the series was aggregated into a monthly series by calculating the quantity for each month in the year using the overall average of each month's daily maximum temperature, as indicated in Figure 4. Also from Figure 1, the daily series appear crowded graphically, making the correlogram difficult to read, therefore the decision to aggregate it into a monthly series can be considered.

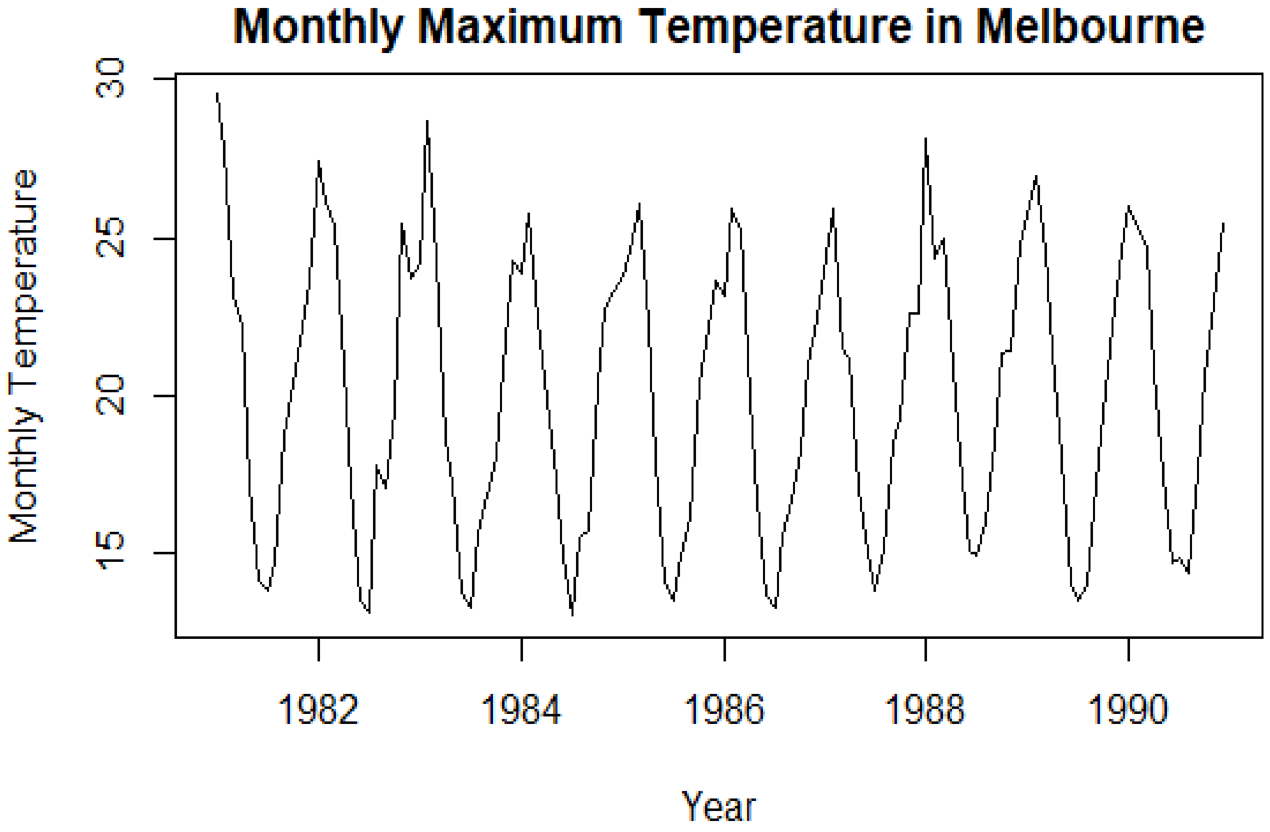

Figure 4: *Time Plot of the Monthly Maximum Temperatures (°C) of Melbourne (1981-1990) reflecting the clear cyclic variation during months.*

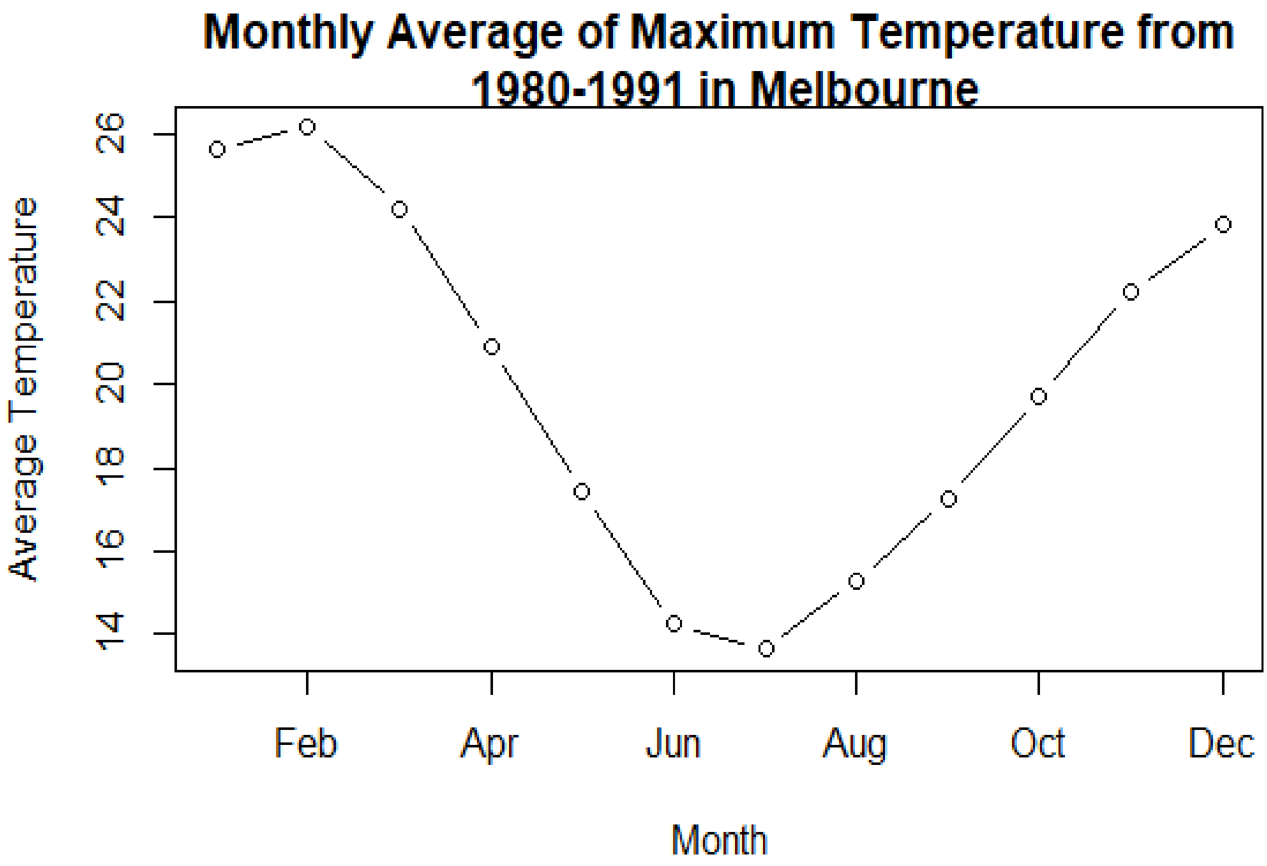

Figure 5: *Plot of Melbourne's Monthly Average Maximum Temperature (°C) indicating Feb and July as the hottest and coldest months resp.*

To study the cyclic variation of the data, the average monthly maximum temperature in Melbourne from 1980 to 1991 was plotted as shown in Figure 5. From figure 5, it can be interpreted that in Melbourne hottest month of the year is usually February and the coldest month is usually July.

Based on the results of all of the previous analyses, it is clear that there is seasonal variation in the data, but no linear trend over time. Furthermore, the ACF plot establishes that the series is not stationary, necessitating differencing. With all of this information in mind, the SARIAMA model will be the best fit for the series.

3 Seasonal Autoregressive Integrated Moving Average (SARIMA)

Autoregressive Integrated Moving Average (ARIMA), is a modeling approach for univariate time series data that excludes seasonal data. For this series, SARIMA (Seasonal Autoregressive Integrated Moving Average) or Seasonal ARIMA is utilized. SARIMA is an ARIMA extension that supports univariate time series data with a seasonal component explicitly.

Selecting hyperparameters for both the trend and seasonal parts of the series is required while configuring a SARIMA. A SARIMA model is denoted by the following notation:

$$\text{SARIMA(p,d,q)(P,D,Q)}_s$$

Mathematically,

$$\Phi_s(B^s)\varphi(B)(1-B)^d(1-B^s)^D y_t = H_s(B^s)h(B)\epsilon_t \tag{1}$$

Where $\Phi_s(B^s)$, and $\varphi(B)$ are autoregressive polynomials of B_s and B of order P and p, and $H_s(B^s)$ and h(B) are moving average polynomial of B_s and B of order Q and q respectively. (1=$B)^d$ and (1=$B_s)^D$ is to allow for non-stationarity arising from seasonal effects and trends. d is the number of times the series is differenced to remove a trend and D is the number of times the series is differenced to remove the seasonal effect. y_t represents the original monthly time series, s stand for the seasonal period and ϵ_t represents the error term.

This report aims to build a model for maximum temperatures in Melbourne and predict the future values. To confirm that the model fitted is the best-fitted one, initially, the data will be divided into training and testing sets, to fit the training data and forecast the testing values and the best model is the one that gives a minimum difference between actual and predicted value. To achieve this, maximum temperatures from January 1981-to December 1989 (monthly averaged series) are considered as a training dataset used to build the model, and temperatures for the year 1990 is considered testing dataset.

Before starting with the SARIMA model, it must be checked whether the series is stationary or not. To detect the seasonality effect in the training dataset sample autocorrelation plot is used, as shown in Figure 6. ACF plot signifies the seasonality present in the dataset with the same cycle length. Therefore, it can be said that the series should be differenced appropriately to remove the seasonality effect.

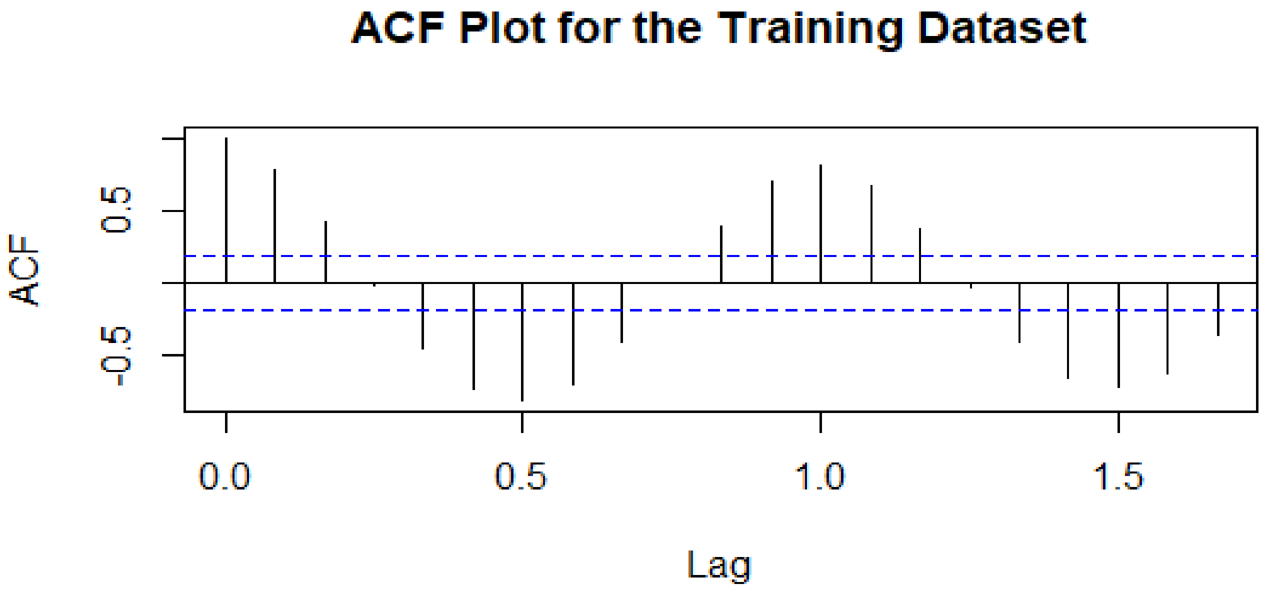

Figure 6: *ACF Plot for the training set confirming the seasonality effect in the data.*

The series was differenced at lag = 12 (since the temperatures are averaged monthly) to eliminate the seasonality variation. Figure 7 depicts the plot of the newly seasonally differentiated series. The series looks to have come to be stationary.

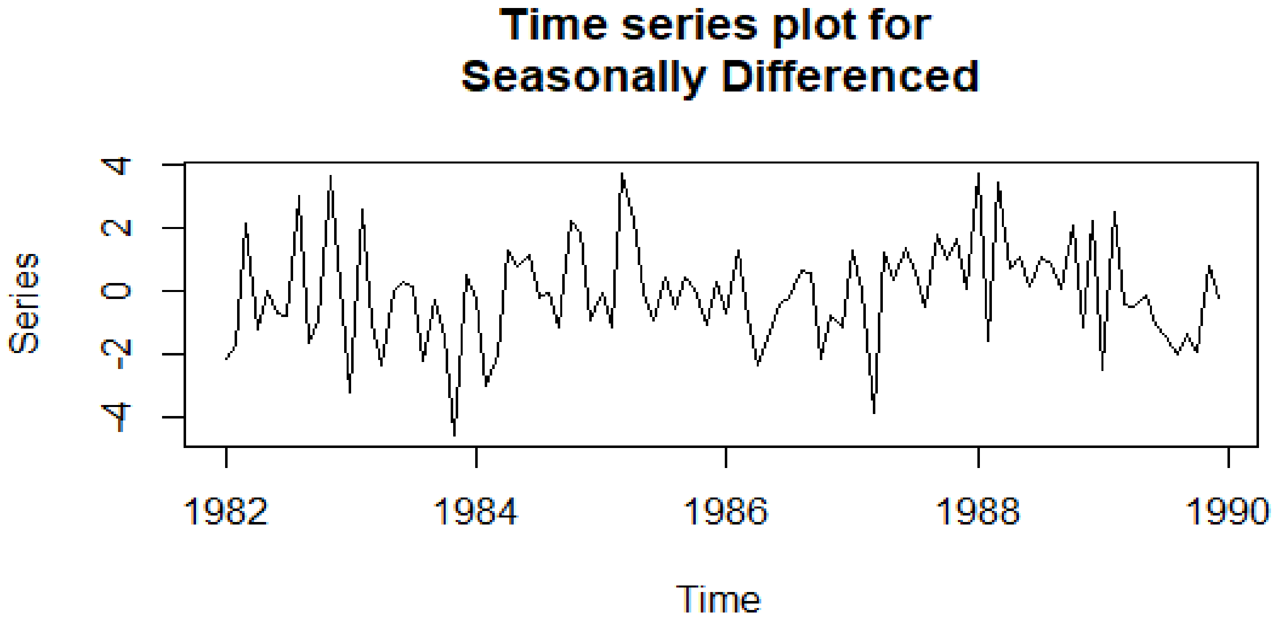

Figure 7: *Differenced training series at lag = 12*

The graph of the ACF and PACF of the seasonally differenced series was produced to get insight into a preliminary order for the ARIMA model to fit the proper SARIMA model.

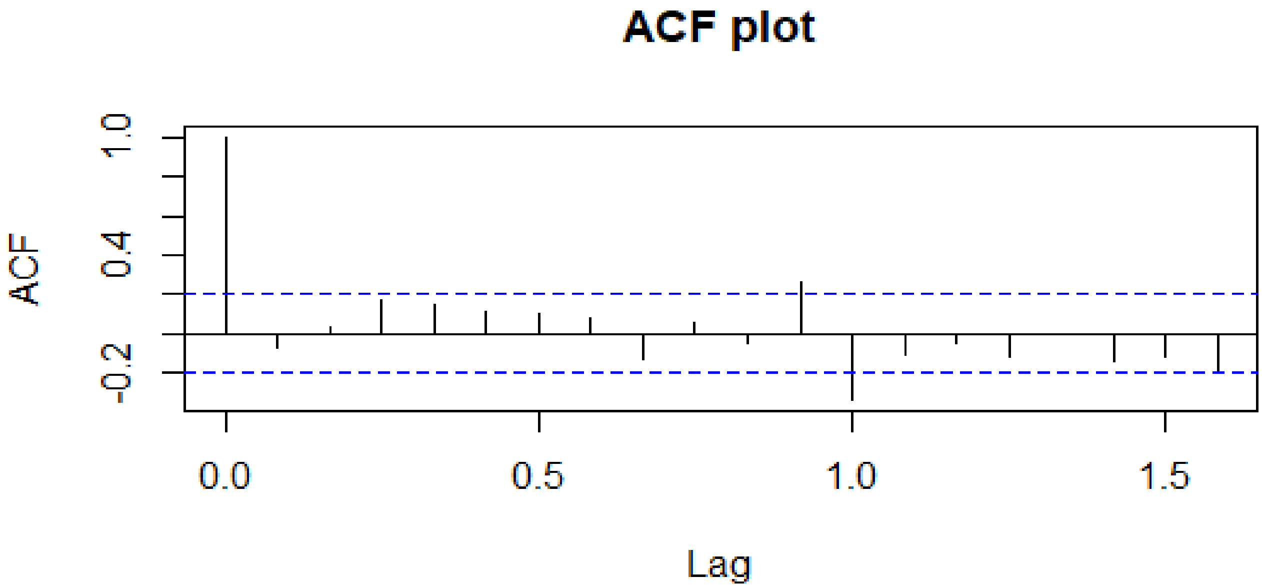

Figure 8: ACF plot the training dataset differenced at lag = 12.

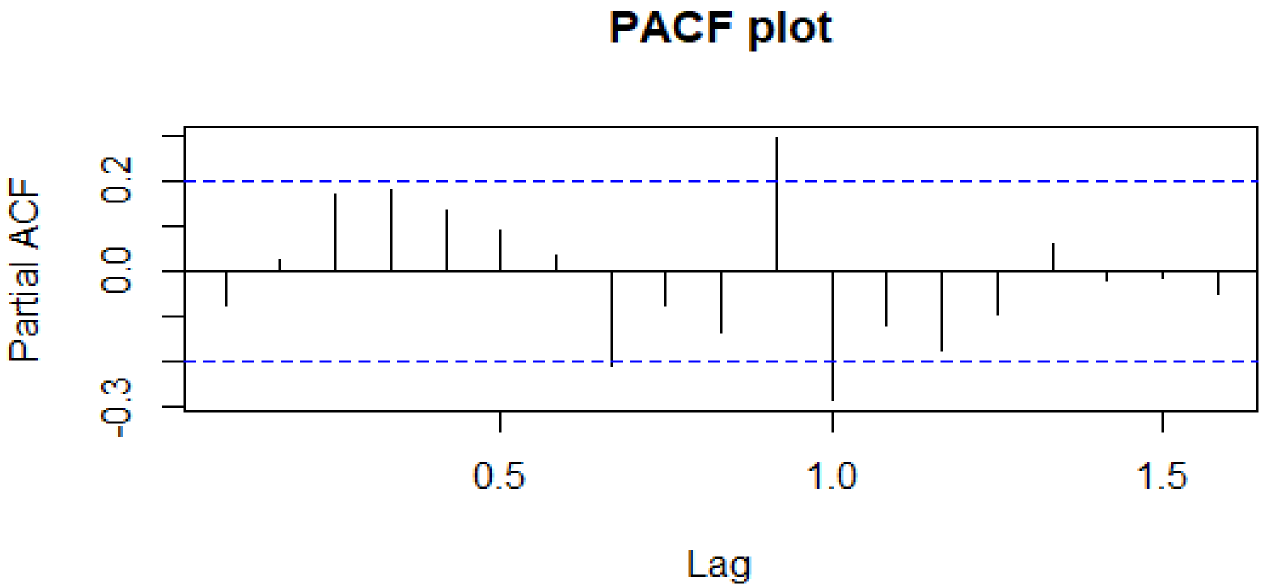

Figure 9: PACF for the training dataset differenced at lag = 12.

The Autocorrelation plot for the differenced temperatures in Melbourne is shown in Figure 8. Even though the ACF eventually tends to zero but spikes are significant at lag = 11 and lag = 12 (the seasonal lag). A partial Autocorrelation plot for the differenced series at lag = 12 is shown in Figure 9. Similarly, in PACF also, even though the values are converging to zero but the spikes are significant at lag = 8,9, and 12. By chance alone, one or two of the autocorrelations for the first few lags could probably be over the 95 percent significance thresholds. To verify this, the Ljung-Box test can be performed in R using *Box.test()* function in *stats* package [2]. It is a statistical test that determines whether any of a collection of autocorrelations in a time series diverge from zero. It is a portmanteau test since it examines ”overall” randomness based on a lot of lags rather than assessing randomness at each individual lag.

H_0: The series is distributed independently (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the series result from randomness of the sampling process).
H_1: The series is not independently distributed; they exhibit serial correlation.

```
Box-Pierce test

X-squared = 0.54359, df = 1, p-value = 0.4609
```

The p-value for the test is 0.46 which is greater than α (=0.05), indicating that the observed correlations in the data are due to the randomness of the sampling procedure.

By visual analysis of the (partial) correlograms, the Box–Jenkins methodology for ARMA models allows one to select the order of an AR(p), MA(q), or ARMA(p, q). But in R, there is a function called *auto.arima()* which tries different values for p,d,q,P,D,Q,s and returns the best fit based on AIC. A model with the minimum AIC is selected as the best-fitted model. Few of the models tried in R, are shown in Table 2. It shows the recommended models and their AIC values after fitting them to the seasonal training set (the best is in bold font for emphasis).

Table 2: *Recommended SARIMA Models and respective AIC values.*

Model	AIC
ARIMA(0,0,0)(0,1,0)[12] (with drift)	371.0198
ARIMA(0,0,0)(2,1,0)[12] (with drift)	343.8846
ARIMA(0,0,0)(2,1,0)[12]	**341.7922**
ARIMA(0,0,0)(2,1,1)[12]	343.7101

SARIMA(0, 0, 0)(2, 1, 0)12 is fitted to the training series was considered to be the best fit and was used to forecast the series based on the findings summarized in Table 2. The SARIMA model fitting was done using the *sarima()* function present in the *astsa* package [3].

Table 3 summarizes the parameters of the aforementioned model. The model has an estimated variance (σ^2) of 1.765 and a log-likelihood of -167.72.

Table 3: *Parameters of the fitted Model.*

Type	Coefficient	Standard Error
SMA(1)	-0.5412	0.0971
SMA(2)	-0.4690	0.1019
Mean	-0.0018	0.0062

Diagnostic checks are performed to ensure that the fitted model is adequate for modeling the series. The residuals of the model reveal no structure, as seen in Figure 10. The graph of the standardized residuals shows that the residuals are standard normal with a mean of 0 and a variance of 1. At various lags, the residuals are not associated, as shown by the ACF Plot. This is also supported by a plot of the Ljung-Box p-values, which shows that all of the p-values are more than 0.05 (points are all above the blue dotted line), indicating that there is no significant connection between the residuals at various lags. The normality of the residuals is also supported by the Normal Q-Q plot of the residuals. Based on all the evidence shown in figure 10, it can be concluded that the fitted model is adequate to structure and forecast the maximum temperatures in Melbourne.

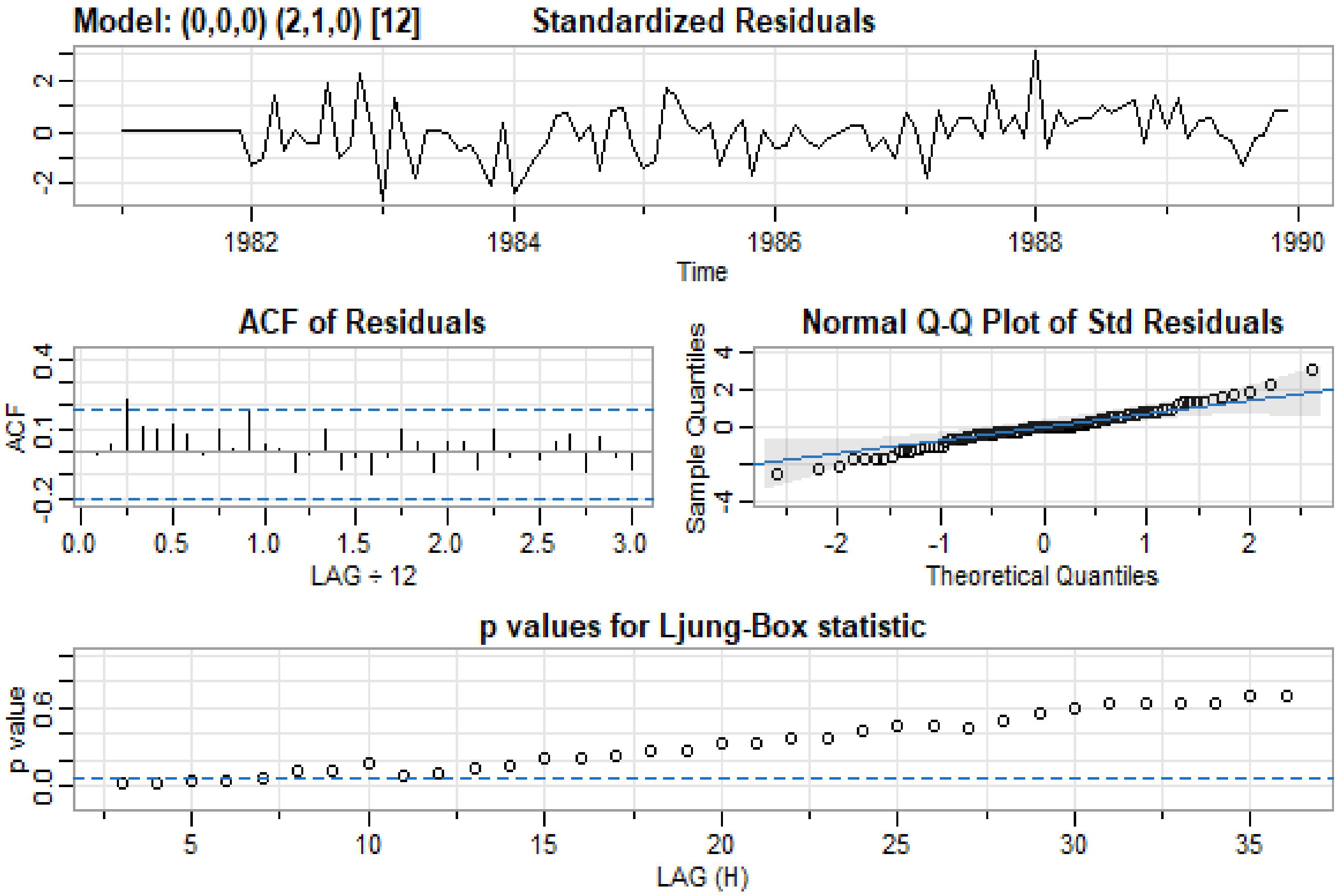

Figure 10: *Graphical Adequacy Checks for the selected model.*

4 Forecasting

The initial monthly series was split into two parts: a train set and a test set. The test set included data from January 1990 to December 1990, while the train set had observations from January 1981 to December 1989. The observations from January 1990 to December 1990 were projected as if they had never been observed to assess the fitted model's predicting performance. The *astsa* library's *sarima.for()* function was used to do this. Forecasting was done with the train set values as they were before to differencing. This is done to make comparisons easier.

The code looked like this:

***sarima.for*(training, 12, p = 0, d= 0, q = 0, P = 2, D = 1, Q = 0, S = 12)**,

where training is the train set before seasonal differencing, the first 12 is the number of step-ahead forecasts to be made based on the observed series (the train set), and the second 12 is the number of step-ahead forecasts to be made based on the observed series (the train set).

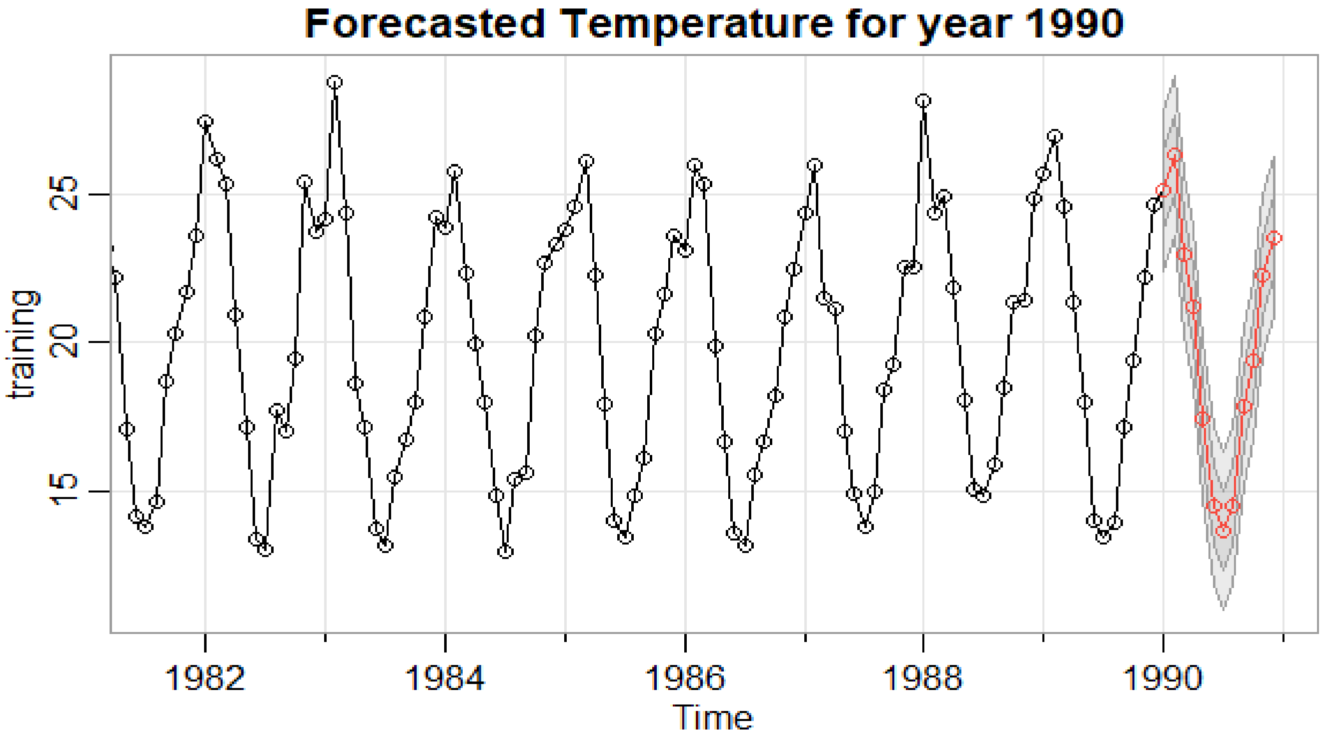

Figure 11: *Graphical representation of Forecasted Values of the Monthly Average Maximum Temperature (°C) of Melbourne (January 1990 - December 1990) with the Prediction interval at 95% and 80% shaded in grey.*

Figure 11 depicts the model's forecasted Average Maximum Temperature (°C) for Melbourne (January 1990 - December 1990) in red with the Prediction interval at 95% and 80% shaded in grey, while Figure 12 is a graph depicting the actual values of Melbourne's average monthly maximum temperature (°C) (i.e., the test set, which includes observations from January 1990 to December 1990) and the values forecasted by the model.

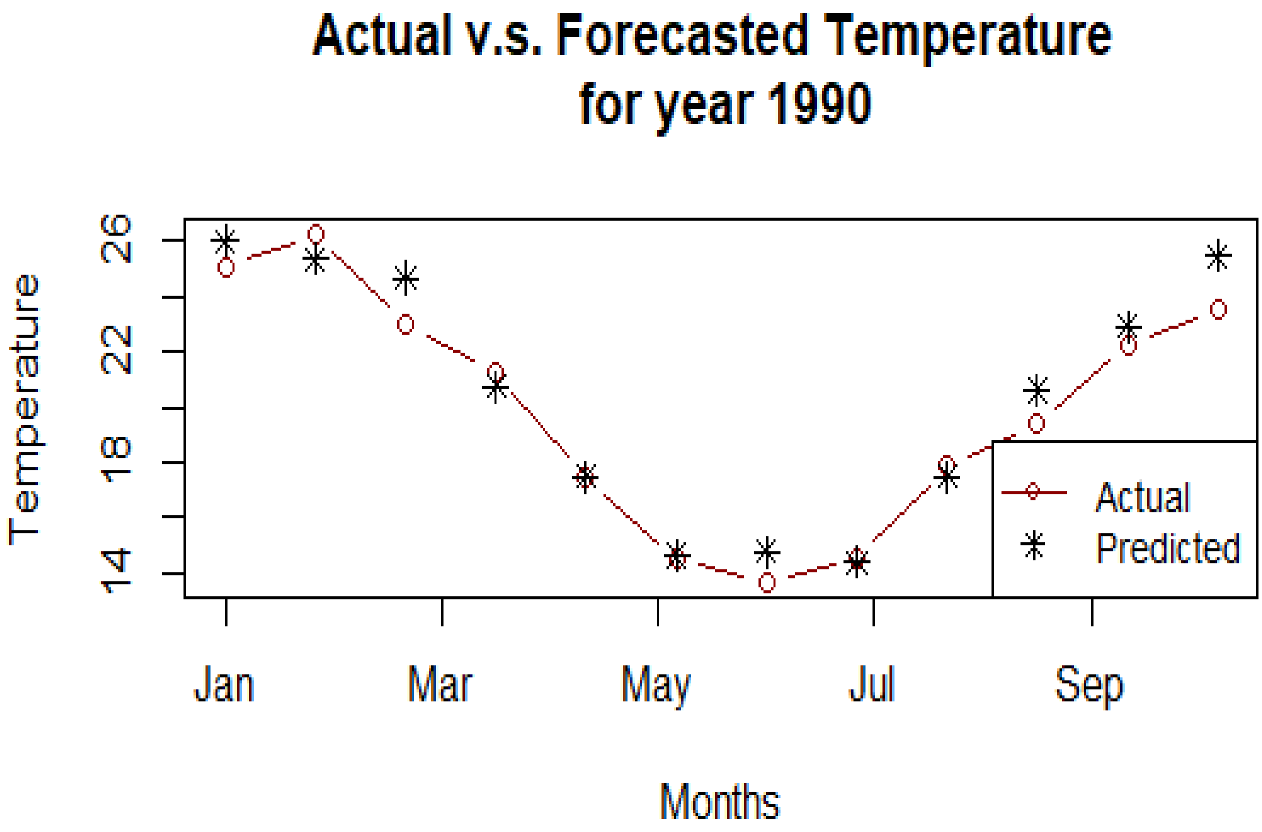

Figure 12: *Graph of Actual and Forecasted Values of the Monthly Average Maximum Temperature (°C) of Melbourne (January 1990 - December 1990).*

Figure 12 demonstrates that actual and predicted temperatures in Melbourne for the year 1990 are pretty close, demonstrating that the model is capable of anticipating the series.

Numeric indicators such as Mean Average Error (MAE), Residual Sum of Squares (RSS), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used to evaluate the fitted model's predicting ability. The fitted model's indicators are summarized in Table 4. These values are small, indicating that the fitted SARIMA model is adequate.

Table 4: *Numeric Forecasting Performance Indicators for the fitted Model.*

Indicator	Value
Mean Average Error (MAE)	0.794
Residual Sum of Squares (RSS)	11.699
Mean Squared Error (MSE)	0.975
Root Mean Squared Error (RMSE)	0.987

Beyond the graphical representation shown in Figure 11, it was also essential to conduct a statistical test to examine if the differences between the actual values in the test set and the values predicted by the fitted model were significant. A two sample t-test was used to accomplish this. However, before doing the two-sample t-test, it was required to ensure that the two groups (actual and predicted values) met the assumption of normality and that their variances were identical. The Shapiro-Wilk test for normality was employed to determine whether the two groups were normal.

H_0: Group's values are normally distributed
H_1: VAlues are not normally distributed

Table 5: *Shapiro-Wilk Test*

Data	Test Statistics	p-value
Actual Temperature	0.885	0.1018
Predicted Temperature	0.935	0.4916

In Table 5, since none of the p-values are less than α (= 0.05), the values in both data sets are assumed to be normal.

The F-test was also employed to determine whether the variances in both data sets were equal.

H_0: Ratio of variances of the two groups equal unity
H_1: The ratio of variances of the two groups is not equal to the unity

Table 6: F-Test

Ratio of Variance	num.df	denom.df	p-value
0.91345	11	11	0.8833

Table 6 summarizes the results of the experiment. Given that the p-value of the result is 0.8833 which is greater than α (= 0.05), it is concluded that both data sets can be assumed to have equal variances.

Now that both of the t-test requirements have been met, it can be checked if there is a significant difference between the actual and forecasted temperatures for the year 1990.

$H_0 : \bar{x}_{Actual} = \bar{x}_{Forecasted}$
$H_1 : \bar{x}_{Actual} \neq \bar{x}_{Forecasted}$

Table 7: *Two sample t-test*

$\bar{x}_{Actual}$	$\bar{x}_{Forecasted}$	Test Statistics	df	p-value
20.369	19.893	0.263	22	0.795

A t-test was conducted and summarized in table 7. The null hypothesis is maintained because the test yielded a p-value of 0.795, which is greater than α $(= 0.05)$, therefore it is determined that there is no statistically significant difference between the actual temperatures and forecasted temperatures for the year 1990 in Melbourne.

5 Conclusion

• Because the maximum daily temperature in Melbourne followed a cyclic pattern from January 1981 to December 1990, the seasonality was removed by differencing the series. To accomplish so, monthly averages of daily maximum temperatures were calculated for each year.

• SARIMA(0,0,0)(2,1,0)12 model was fitted to the training dataset (January 1981 to December 1989).

• The maximum averaged monthly temperature was anticipated from January 1990 to December 1990 using information from the fitted model and afterward compared to the actual temperature over that period.

• The fitted model's prediction ability was assessed using several indicators such as Mean Average Error (MAE), Residual Sum of Squares (RSS), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These values were low, indicating that the SARIMA model that was fitted is sufficient.

• A two-sample t-test was used to determine if the fitted model is the most accurate. The t-test revealed that the discrepancy between the anticipated and actual temperatures for 1990 was not significant, demonstrating that the model is acceptable.

References

[1] Selva Prabhakaran. *Time Series Analysis with R.* URL: `http://r-statistics.co/Time-Series-Analysis-With-R.html`. (accessed: 21.04.2022).

[2] R Core Team. *R: A Language and Environment for Statistical Computing.* ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: `http://www.R-project.org/`.

[3] David Stoffer. *astsa: Applied Statistical Time Series Analysis.* R package version 1.14. 2021. URL: `https://CRAN.R-project.org/package=astsa`.

[4] Dr Kostas Triantafyllopoulos. *MAS61005: Time Series.* University of Sheffield. 2021.