Daily Temperatures in Melbourne: time series analysis and forecast

Q Yun

2025-05-11

1. Introduction

Analysing weather data and making reasonably accurate predictions can not only bring huge practical benefits for human economic activities, but also help us to understand the long-term climate change pattern. Auto Regression Integrated Moving Average (ARIMA) has been widely used in projects on weather data modelling (eg. Dahiya, 2024), and is also to be used in this project. The dataset in the project contains the daily maximum temperatures (degrees Celsius) in Melbourne of Australia over a period of 10 years (1981 and 1990). R will be the tool used in the analysis and modelling of data in this project.

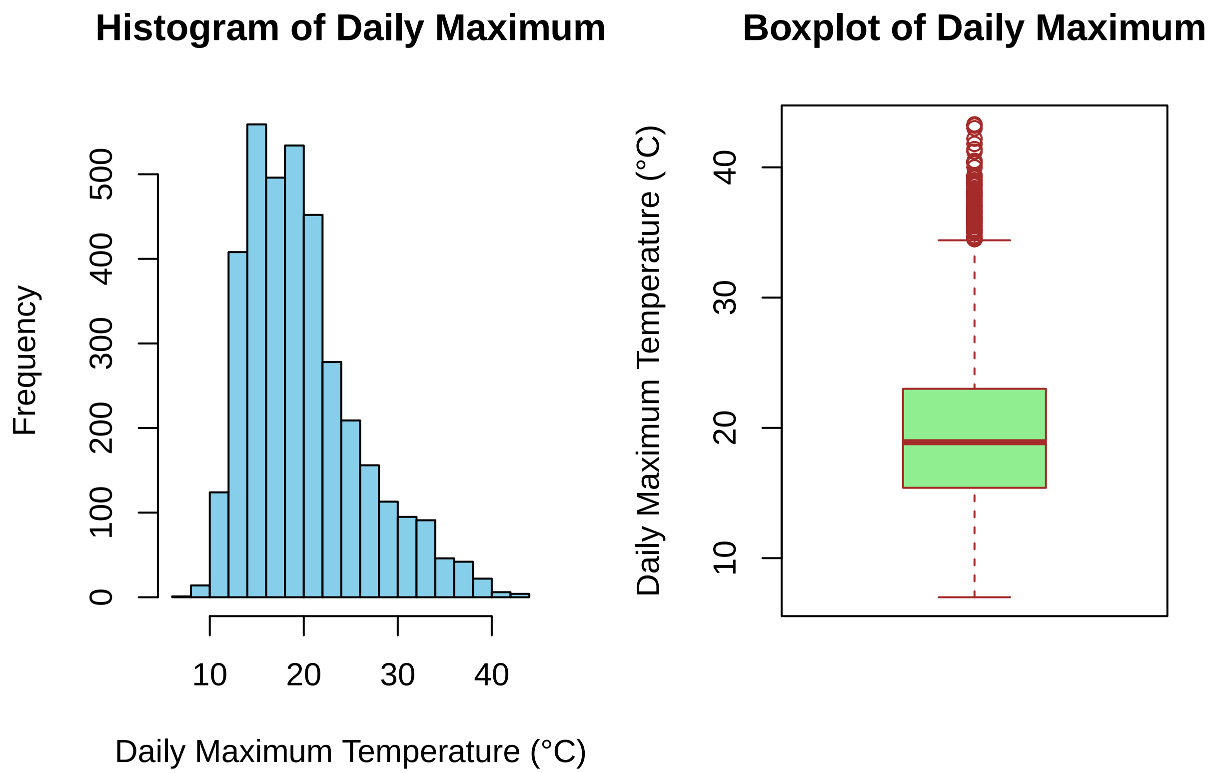
There are a couple of objectives in this project:

- evaluate the possible long-term trend of the temperature in Melbourne;
- understand the seasonality;
- apply an appropriate time series model based on our understanding of the trend and seasonality;
- forecast the temperatures and check the model performance against a test set.

2. Data exploration

Although we are going to use the ARIMA approach for the dataset, it is still essential to carry out data exploration in order to identify the shape, patterns or possible outliers of the data. An examination of the dataset confirms that it does not contain any missing data, which simplies our analysis.

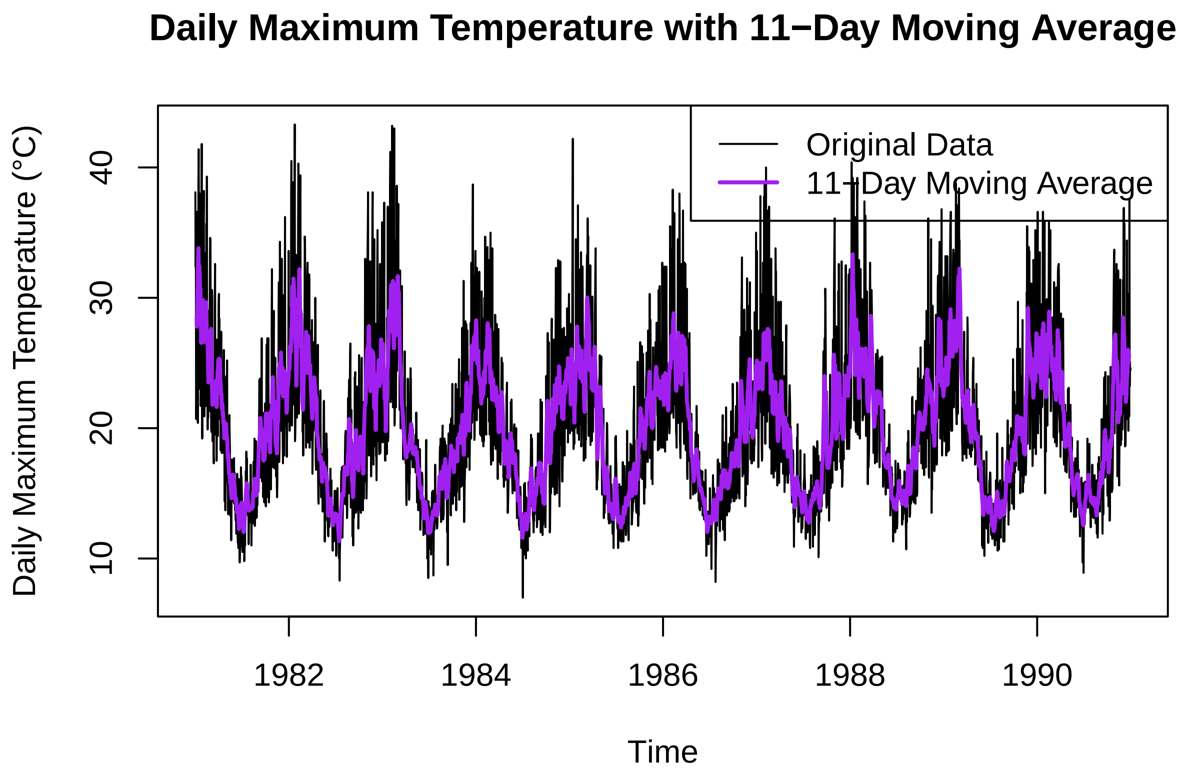
After plotting a histogram of the daily maximum temperatures, it can be seen that there is only one mode in the data. The distribution of the data roughly resembles a bell shape, although with a significant skewness to the right. A box plot of the temperatures reveals quite a few possibly outliers of unusually high temperature, which is inline with the right-skewed distribution.



The summary statistics of the maximum temperatures over the 10 years are as follows,

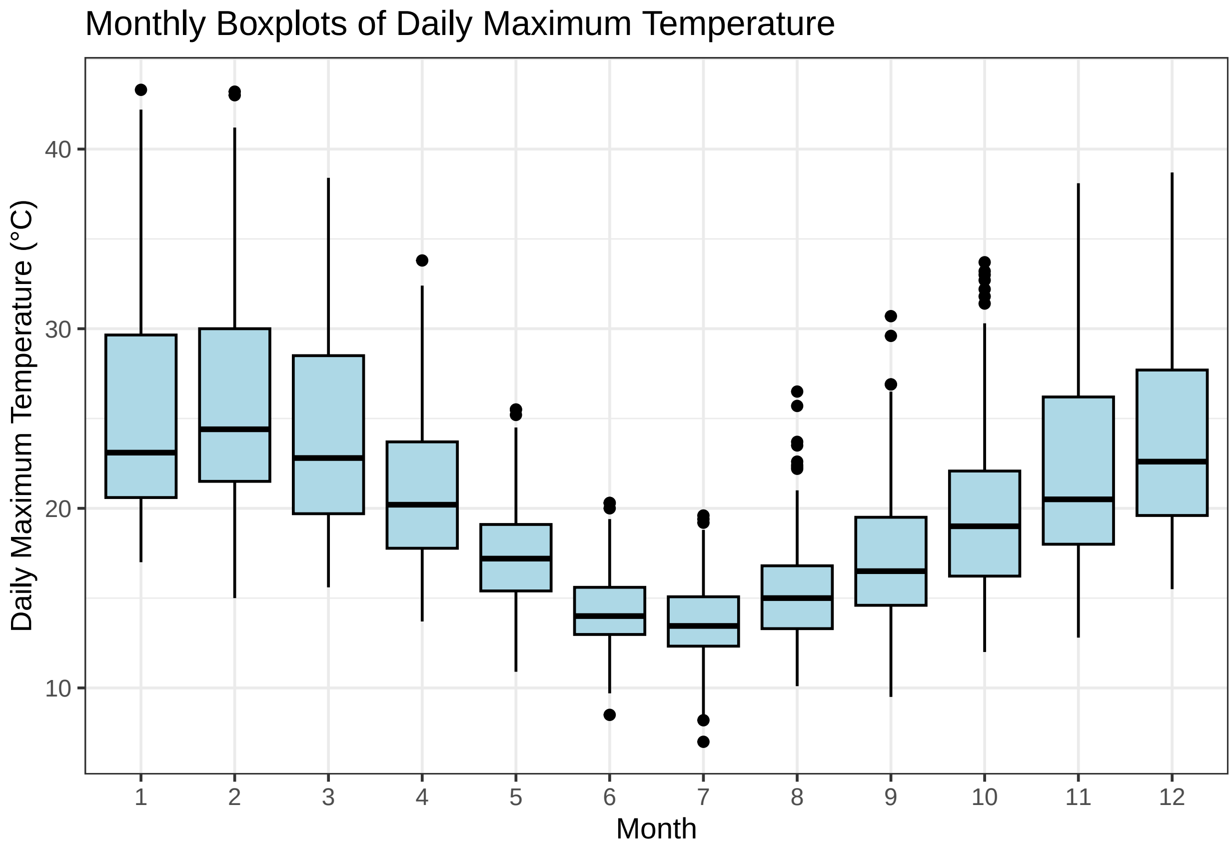
	Summary Statistics
Min.	7.00000
X1st.Qu.	15.40000
Median	18.90000
Mean	20.00915
X3rd.Qu.	23.00000
Max.	43.30000

In order to manipulate the data more effectively, it is also essential to transform the date information in our dataset to the Date format in R. After the transformation of data format, a time series of the daily maximum temperature over the 10-year period is plotted. In order to reduce the noise of daily data, a 11-day moving average of the daily maximum temperature is superimposed onto the plot for a better understanding of the trend.



It can be seen clearly from the above plot that there is a typical periodic cycle in our data, which is understandable due to the annual pattern of our weather. Fortunately, there doesn't seem to be any obvious increasing or decreasing trend. However, if we look at the 11-day moving average, the troughs of lowest temperatures stay rather stable, while there seem to be a small wave for the peaks over the period. This may suggest a cyclical pattern for the high temperatures in the summer, which means a couple of consecutive hot summers may be followed by a couple slightly mild summers. This wave is not very significant, therefore we may choose to leave it if our ARIMA model works fine.

There may also be patterns for the temperatures in different months. Therefore we aggregate the temperatures fro each month over the 10-year period, and produce box plots for different months. It reveals that the range and variability of the temperatures in the winter months (June and July in particular) are much smaller, while those in the summer (January and February in particular) are much larger, which indicates that extreme temperatures often happens in the summer.



3. Model fitting

3.1 Data transformation and exploration of monthly data

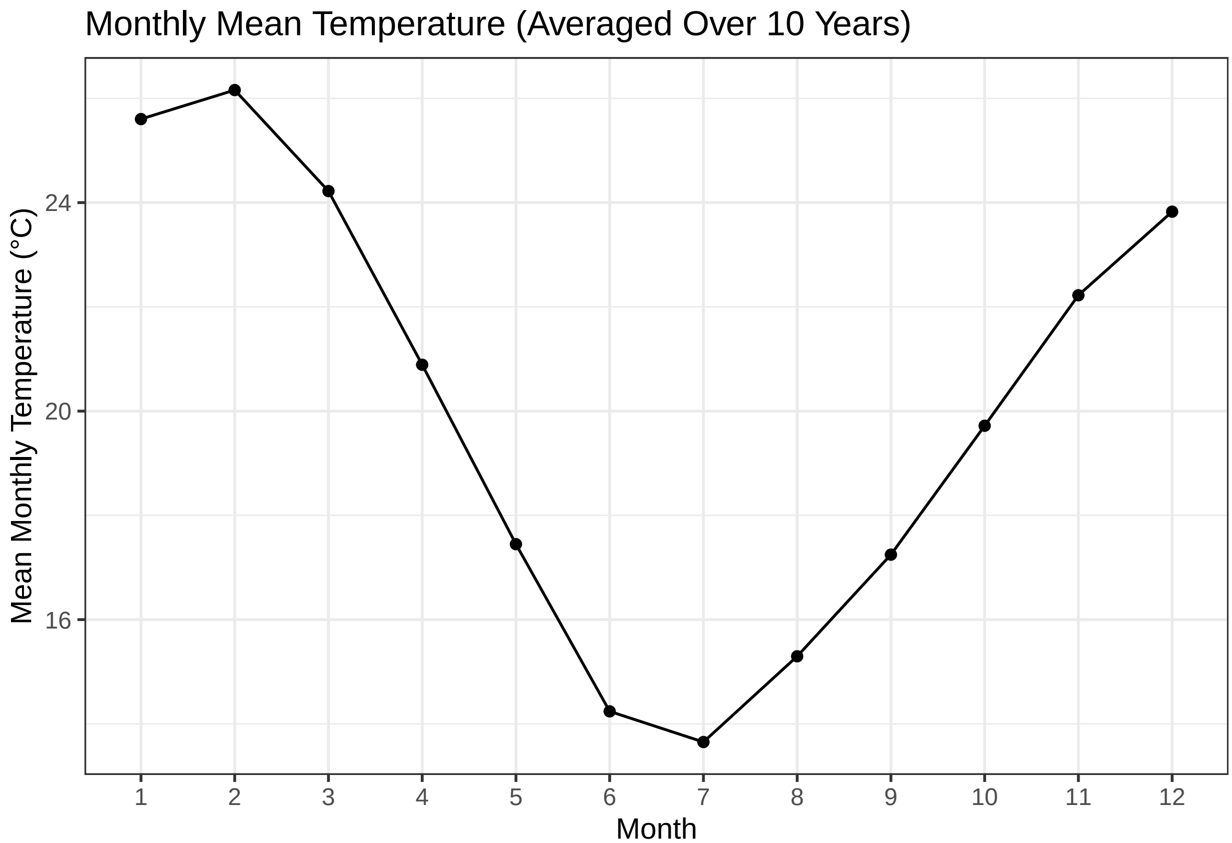
Since initial data exploration has revealed a yearly cycle of the daily maximum temperatures and no specific short-term patterns for the daily temperature is identified, it is appropriate to aggregate the daily maximum temperatures into monthly mean temperatures. As we are more interested in long-term patterns rather than the change of daily temperatures, this aggregation will help to reduce the impact of the outliers and the noise of daily temperatures, hence facilitate our analysis of the seasonality.

Table 2: Monthly mean maximum (first 10 months)

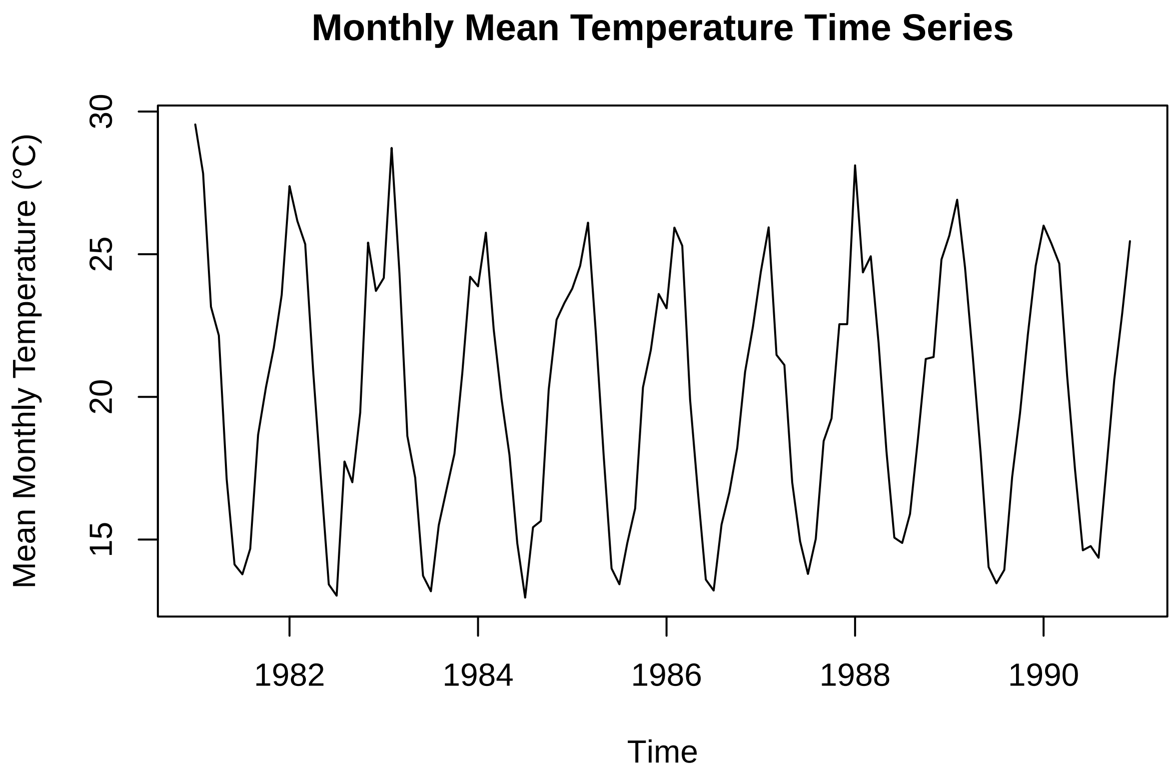
Month	Year	Monthly mean maximum (°C)
1	1981	29.54839
2	1981	27.83214
3	1981	23.15806
4	1981	22.15333
5	1981	17.11613
6	1981	14.12667
7	1981	13.78065
8	1981	14.67742

Month	Year	Monthly mean maximum (°C)
9	1981	18.67000
10	1981	20.33548

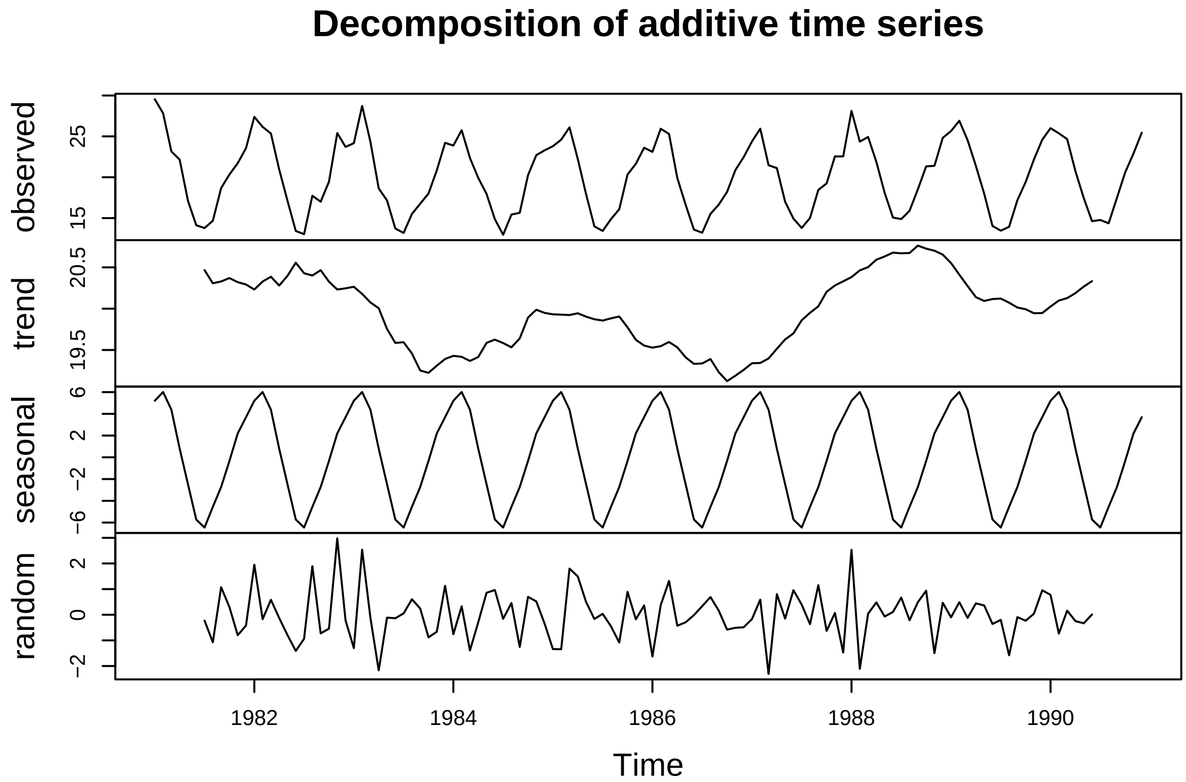
Although it is common knowledge that temperatures usually follow an annual pattern, the monthly temperatures are averaged over the 10-year period and then plotted. No quarterly or semi-annual pattern is found in the following plot.



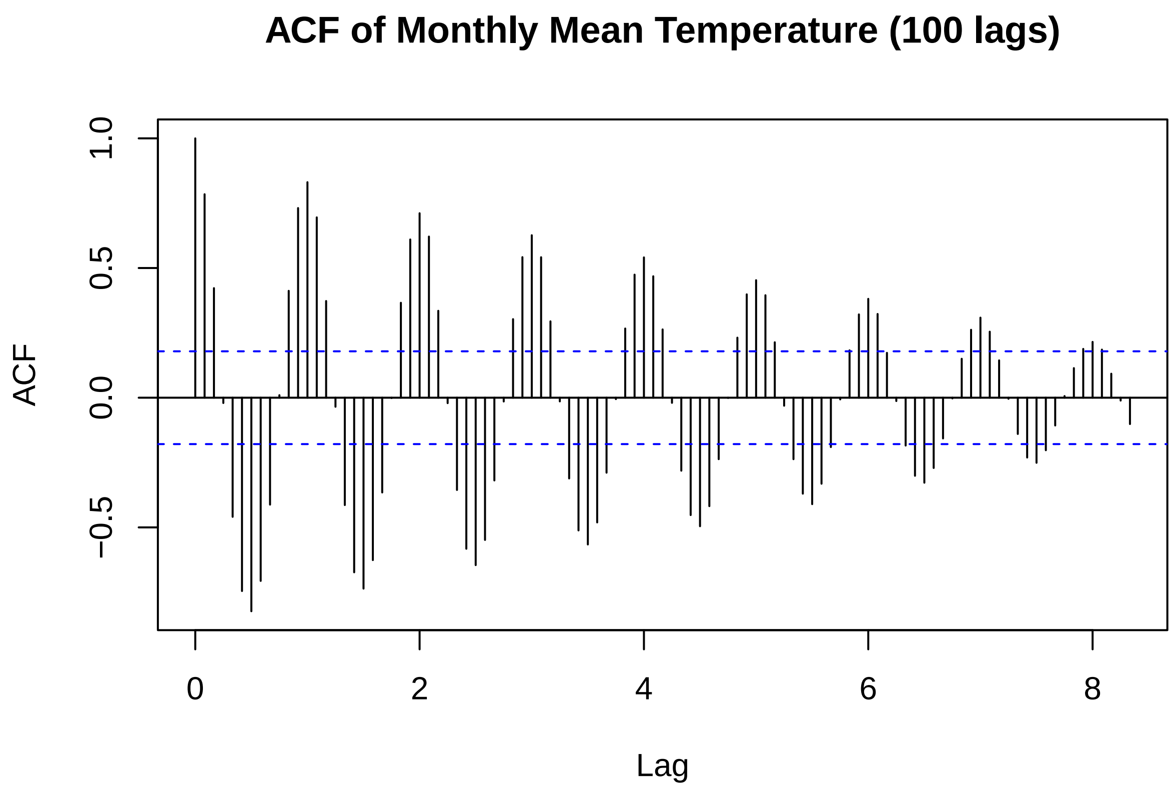
The monthly mean temperatures are then plotted over the whole 10-year period. With less data points than the daily temperatures, the monthly data displays a clearer annual cycle.



The monthly data is then decomposed by applying the decompose function (with the additive model). After the docomposition, it can be seen that the seasonals in the plot appear to be rather constant and persistent. In addition, it also becomes more evident that there is a waving trend with a small range of about 1.5°C, which seems to confirm our earlier observation in the data exploration.



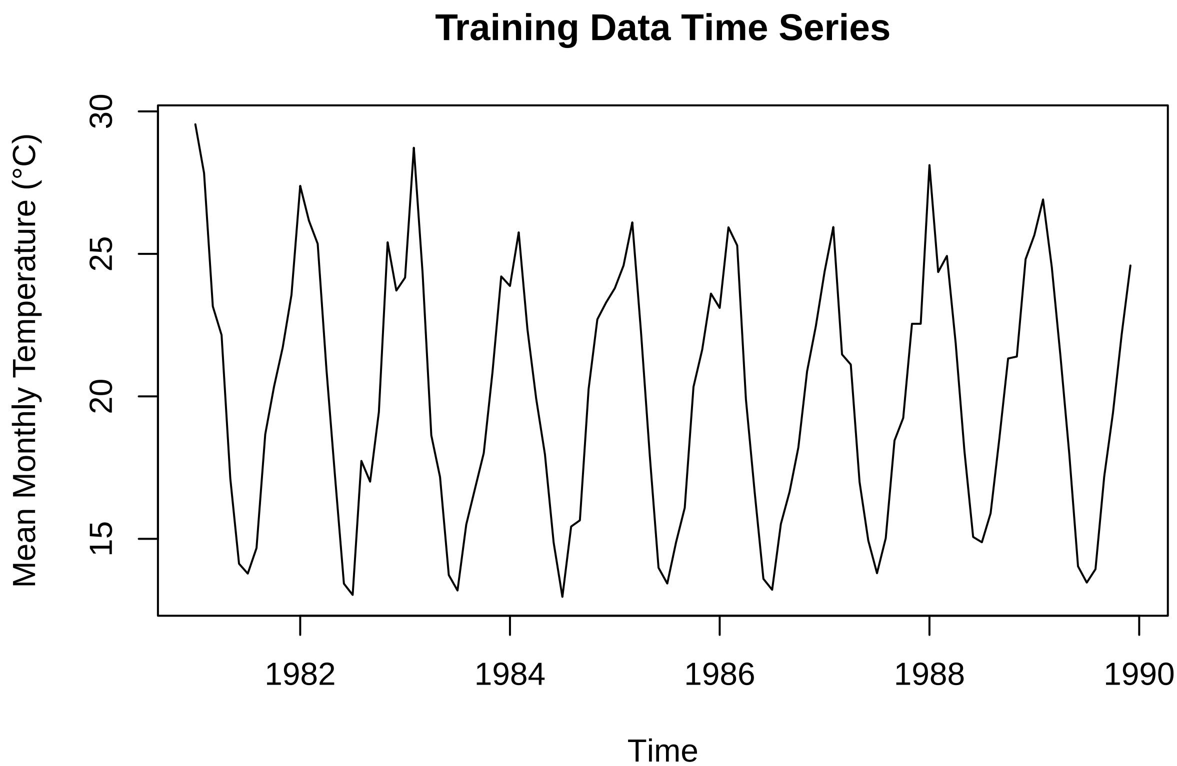
The ACF plots has also confirmed the strong auto-correlations in the data, and the strong seasonality also features prominently in the ACF plot.



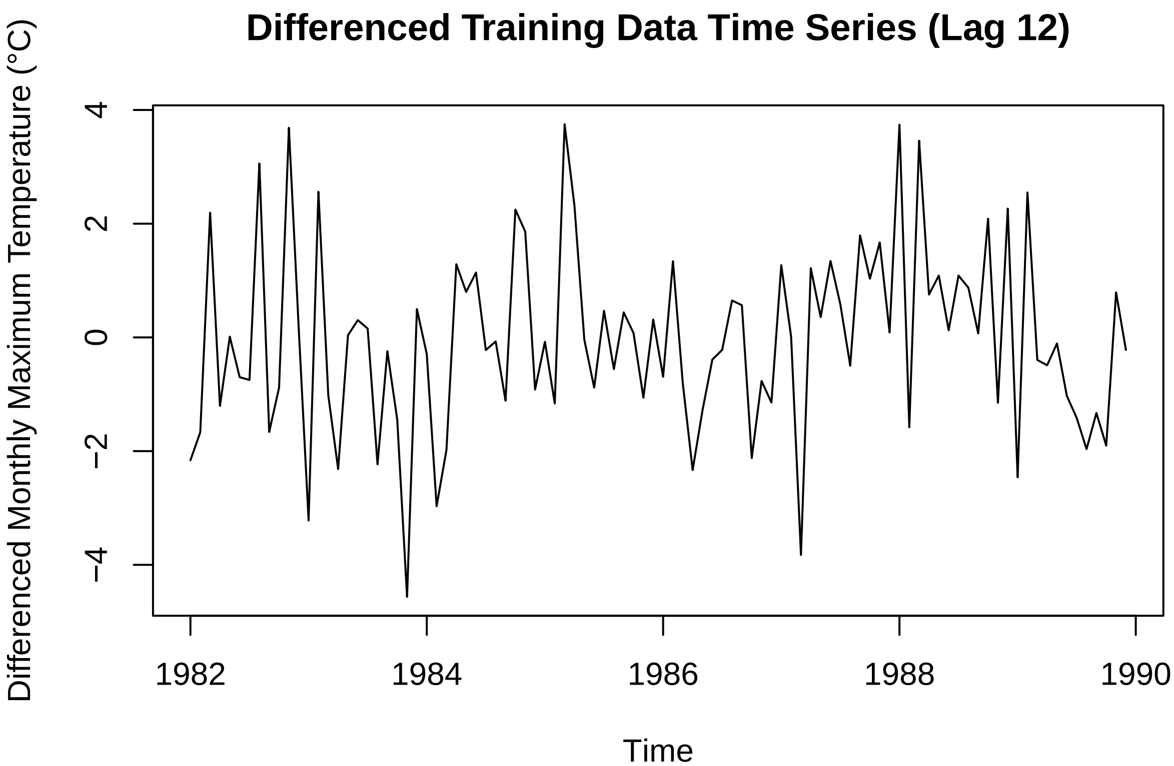
3.2 Training set and its data exploration

In order to assess the performance of our ARIMA model, the 120 data points in the monthly mean data are split into a training and a test set. The training set consists of the data points during the first nine years (the first 108), and the test set includes the last 12 data points for the most recent year of 1990.

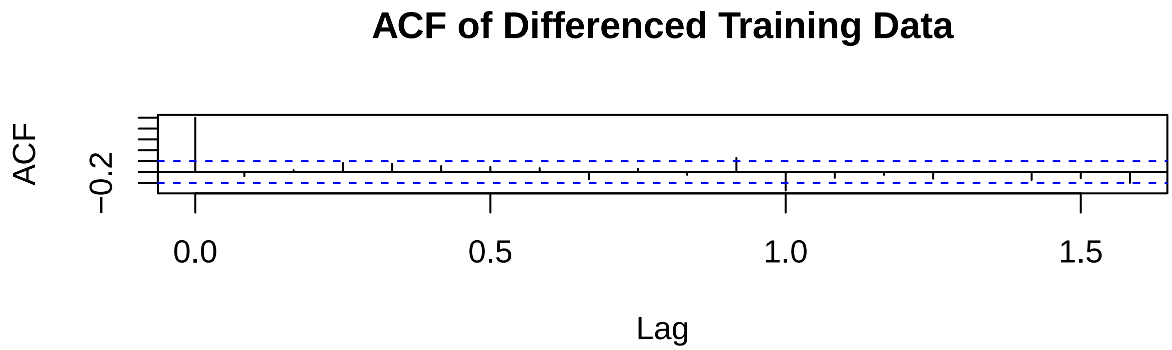
After the split of the datasets, we would like to double check if the pattern in the training set remains unchanged, and this can be confirmed in the following time series plot of the training data.

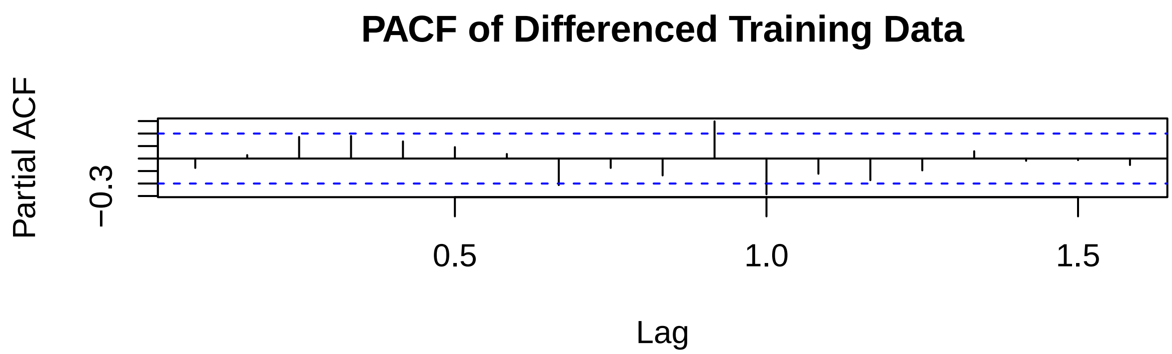


Due to the strong seasonality and a small wave trend present in the data, a seasonal difference with a lag of 12 is probably the first essential step to achieve stationarity. The seasonally differenced training data is then plotted, and it can be seen that the strong seasonality has disappeared.

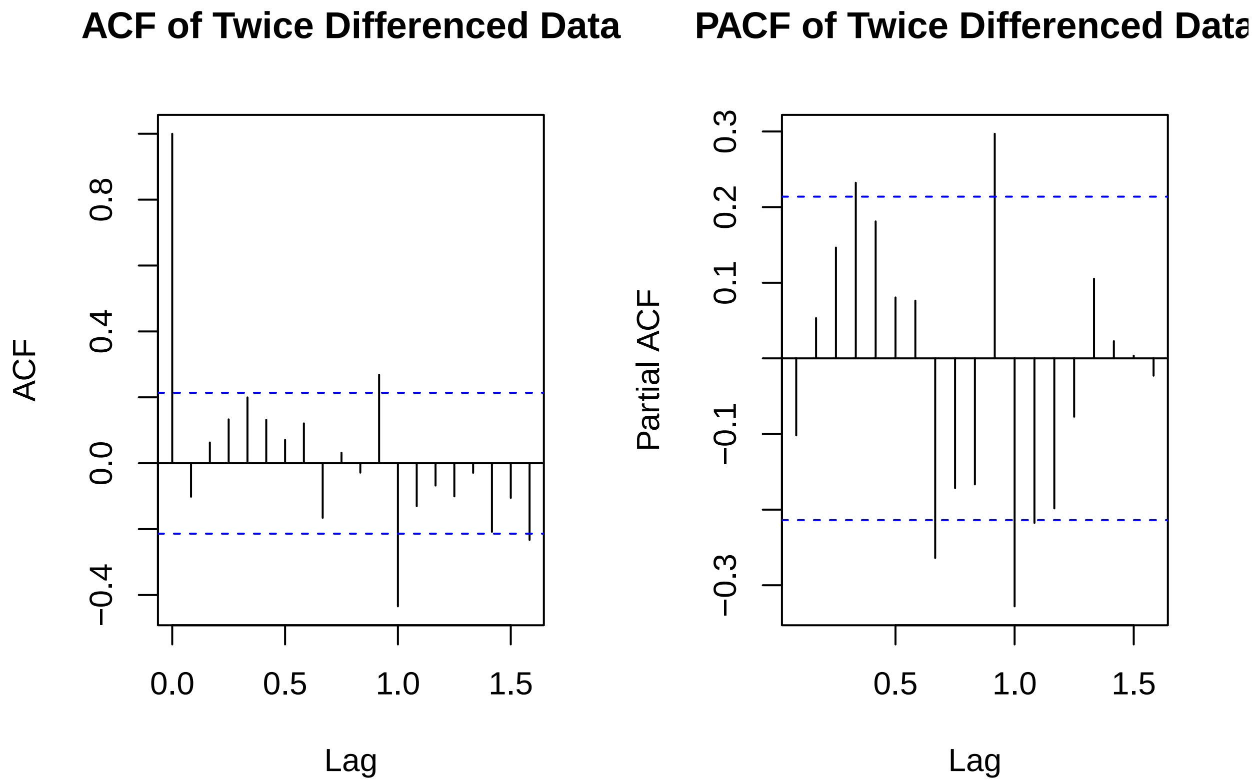


However, we still need to examine the ACF and PACF of the seasonally differenced data to check the auto-correlations.

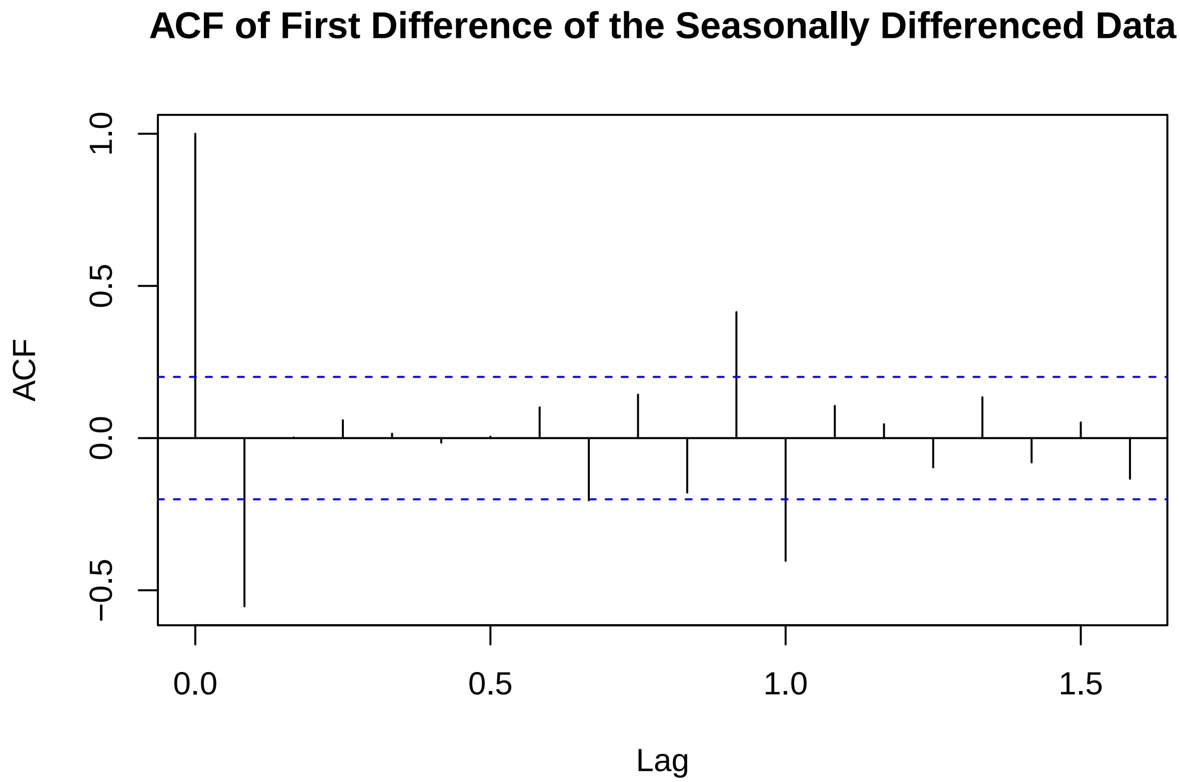


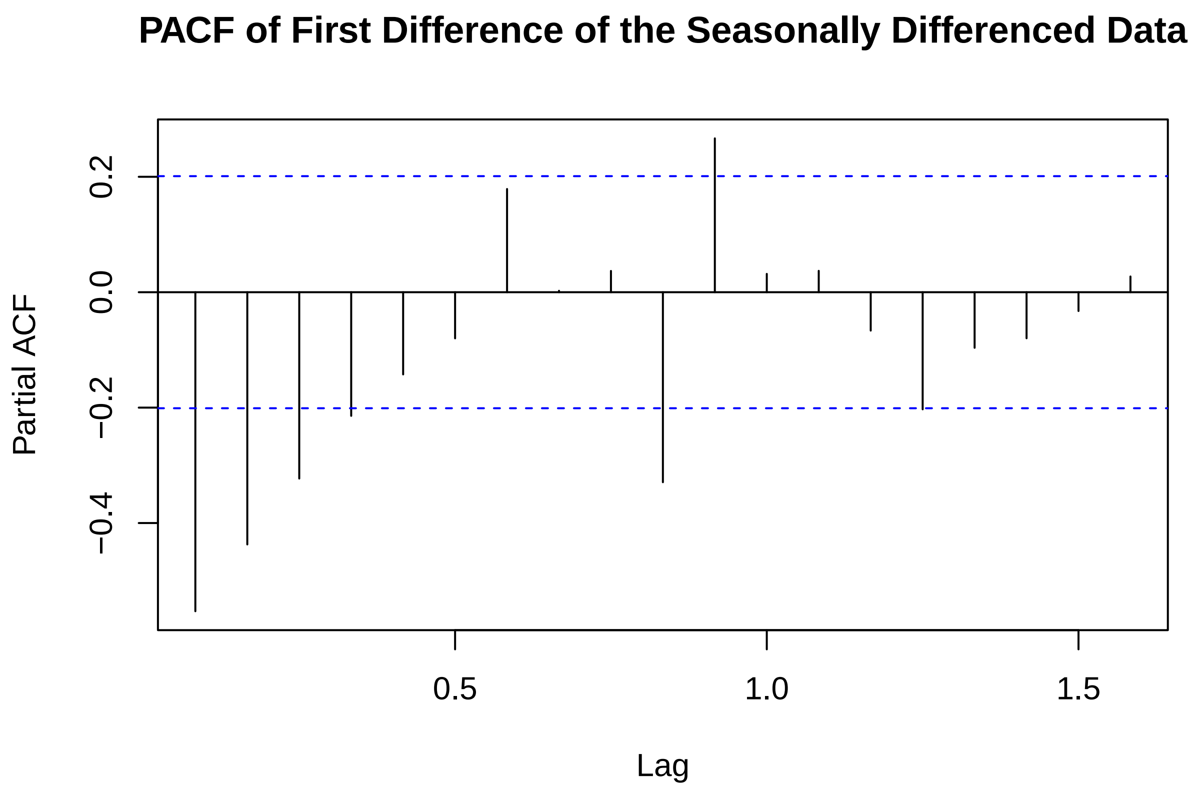


As can be seen from the above two plots, the seasonality has largely been removed. However, there are still two spikes around a full cycle of 12 lags in both plots, and there still appear to be slight seasonality in the PACF plot. A Box-Ljung test with extremely small p-value also indicates the auto-correlations still exist in the seasonally differenced data. A further round of seasonal difference might be an option. However, it turns out that twice seasonal difference does seems to be worse off, particularly in the PACF plots.



A first difference after the seasonal difference may be the other option. However, it can be seen from the following ACF and PACF plots that this option doesn't work either, although the remaining seasonality in the PACF seems to become weaker.


ACF of First Difference of the Seasonally Differenced Data
ACF
1.0
0.5
0.0
−0.5
0.0
0.5
1.0
1.5
Lag

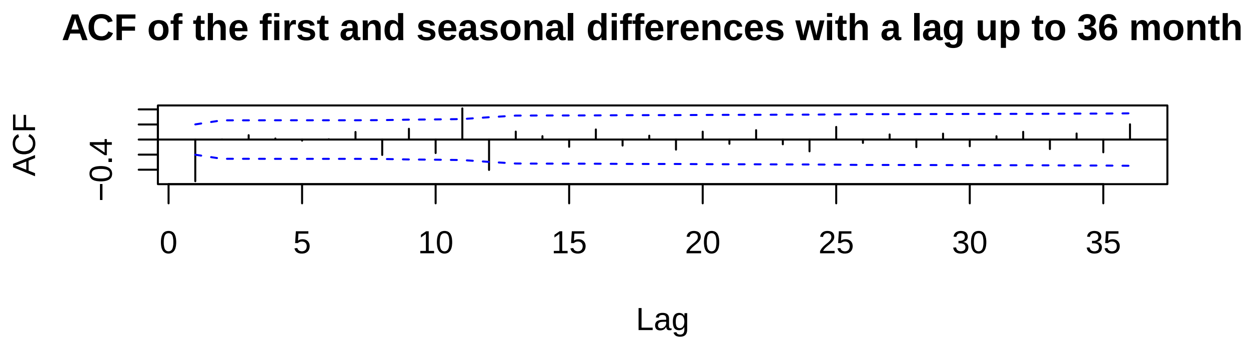


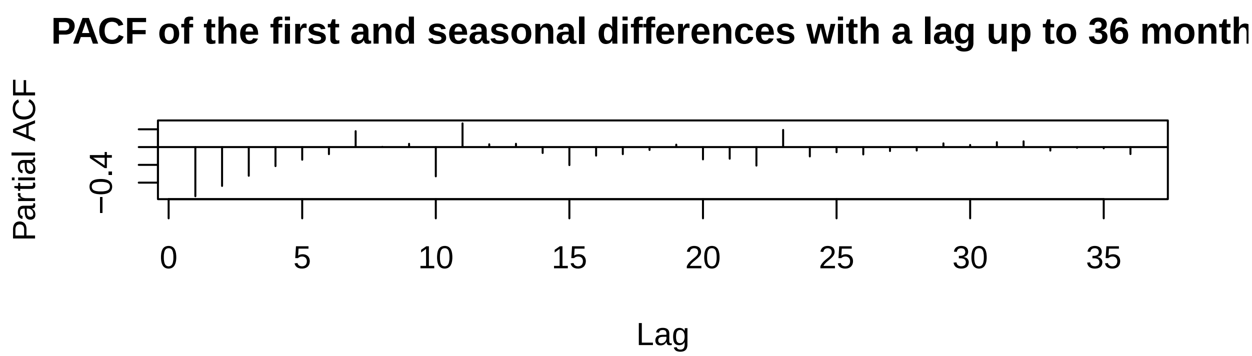
There is a spike at a lag of one month in the above ACF plot, which may suggest an MA(1) term for the differenced data, while the spikes up to a lag of 2 months in the PACF plot may suggests a AR(2) terms. The Extended Autocorrelation Function (EACF) is also helpful in identifying the orders of MA and AR terms. The rows in the EACF represent the AR term and the columns represent MA terms. By looking at the EACF result, it appears that a AR(1) term and an MA(1) or MA(2) might be appropriate. However, by taking into account the principle of parsimony, we will mainly consider an ARIMA model consisting of AR(1) and MA(1) terms.

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x x o x x x x x o x x  x  x  x
## 1 x x o x x x x x o x x  x  x  x
## 2 x o o o o o o x x x x  o  o  o
## 3 x o x o o o o x o o o  o  o  o
## 4 x o x o o o o x o o o  o  o  o
## 5 x o x o o o o x o o o  o  o  o
## 6 x x x o o o o x o o o  o  o  o
## 7 x x o o o o x o o o o  o  o  o
```

Table: EACF Values for Temperature Time Series

We further investigate the two spikes near the 12-month lag in the above plots for the seasonal and then first differenced data. We change the confidence interval type to be ‘ma (moving average)’ for the ACF plot and also plot the PACF up to a lag of 36 months.

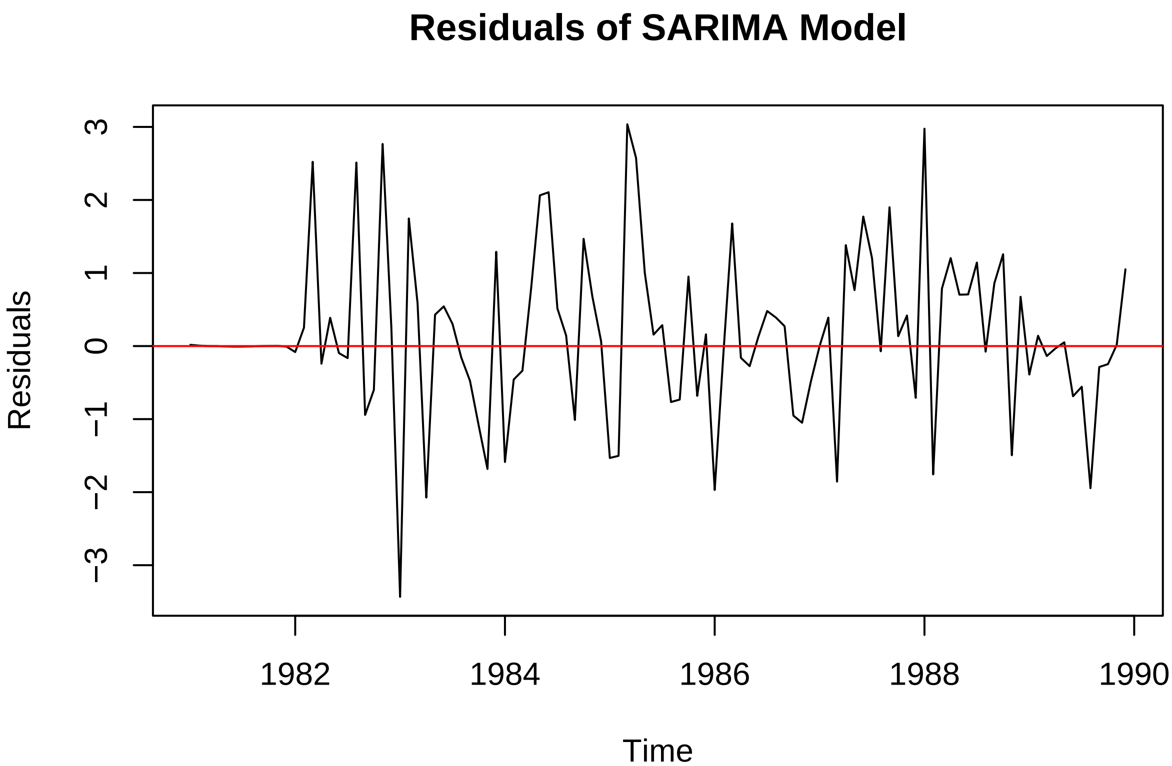




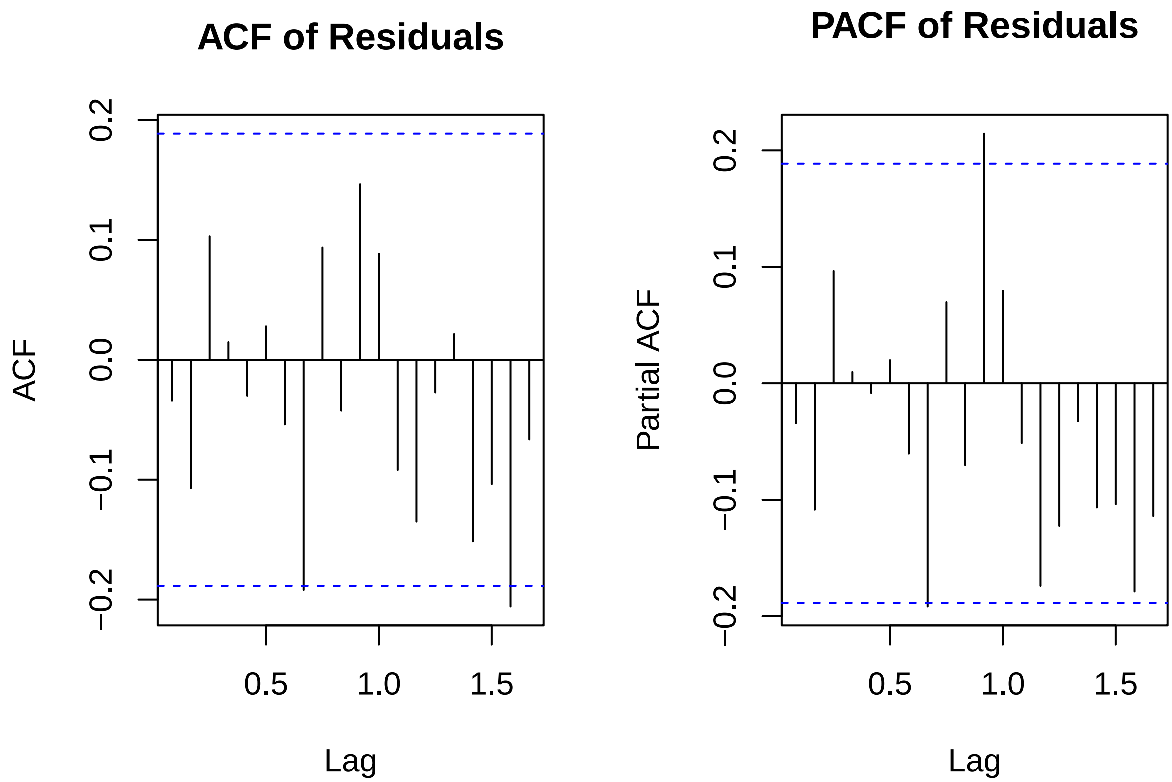
It can be seen that the auto-correlations disappears after lags further than 12 months. Therefore we may consider the two spikes to be the result of chance.

SARIMA models

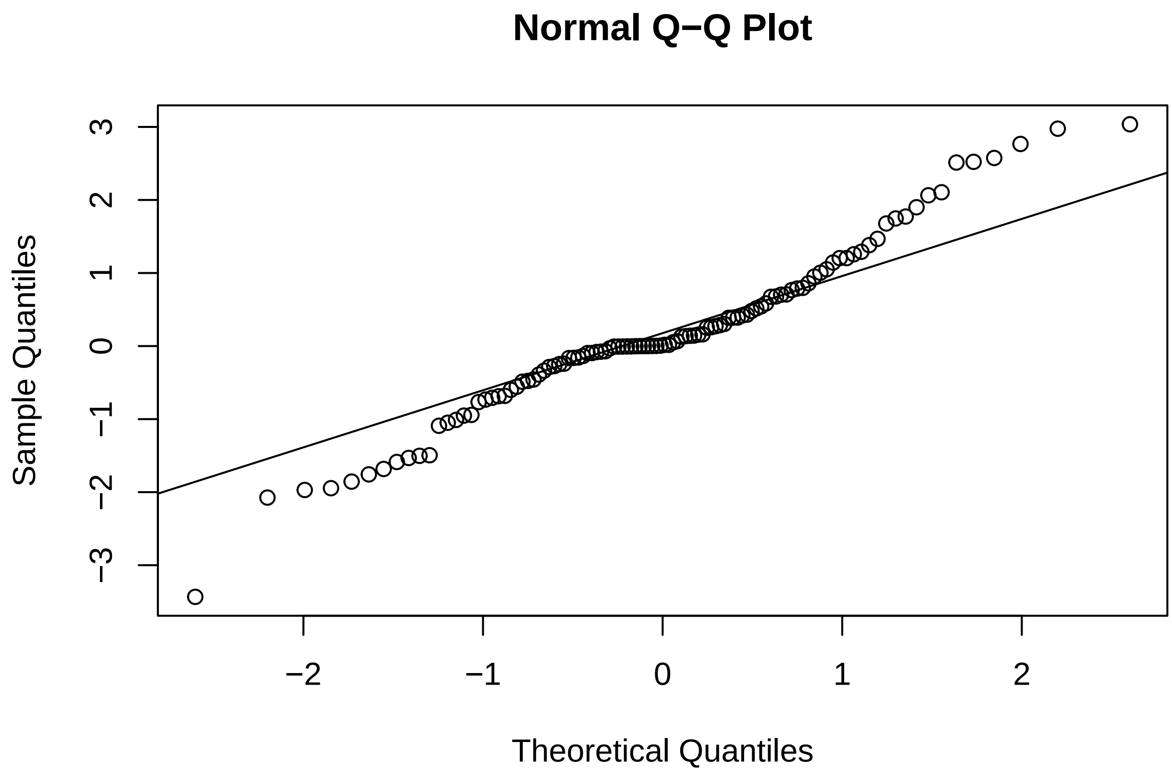
We first apply a $SARIMA(1,1,1) \times (0,1,1)_{12}$ model to the training set, and plot the time series of the residuals of this model.



We then plot the ACF and PACF of the residuals after fitting the model. It can be seen from the following plots that the there is no strong correlations among the residuals. The Ljung-Box test reports a p-value of 0.3826, which suggests a white noise process for the residuals.



The normality of the residuals is also checked by the QQ-plot and the Shapiro-Wilk normality test, and it seems the distribution of the residuals deviates slightly from a normal distribution.



Although this model appears to be a possible option, we would like to compare a few more models with different parameters. Since it seems that another first difference doesn't appear to help much with the model, by taking into account the principle of parsimony, we will only consider the seasonal difference of the data and set the p, q, P, Q parameters to be between 0 and 1. Therefore, different combinations of parameters are applied to the model of $SARIMA(p, 0, q) \times (P, 1, Q)_{12}$.

p	q	P	Q	AIC	SSE	p_value
0	0	0	0	369.0039	257.0960	0.0004789
0	0	0	1	341.5619	143.7588	0.0692462
0	0	1	0	357.2186	218.1229	0.0408876
0	0	1	1	341.8902	146.0367	0.1420646
0	1	0	0	370.5081	255.7590	0.0006899
0	1	0	1	343.3936	143.5062	0.0864796
0	1	1	0	359.1892	218.1266	0.0395413
0	1	1	1	343.8743	145.9067	0.1453514
1	0	0	0	370.4654	255.6429	0.0007373
1	0	0	1	343.3863	143.4951	0.0878858
1	0	1	0	359.1876	218.1271	0.0395388
1	0	1	1	343.8734	145.8991	0.1456588
1	1	0	0	372.4521	255.6069	0.0007821
1	1	0	1	341.1376	138.5541	0.2606685
1	1	1	0	361.1852	218.1250	0.0399212

p	q	P	Q	AIC	SSE	p_value
1	1	1	1	344.2263	139.1250	0.2455983

According to the above results, the model with p=1, q=1, P=0, Q=1 has the lowest AIC value, the lowest SSE. and the highest p-value in the Ljung-Box test, therefore we will choose $SARIMA(1,0,1) \times (0,1,1)_{12}$ to be the model. We then fit this model to the training data and find the coeffients.

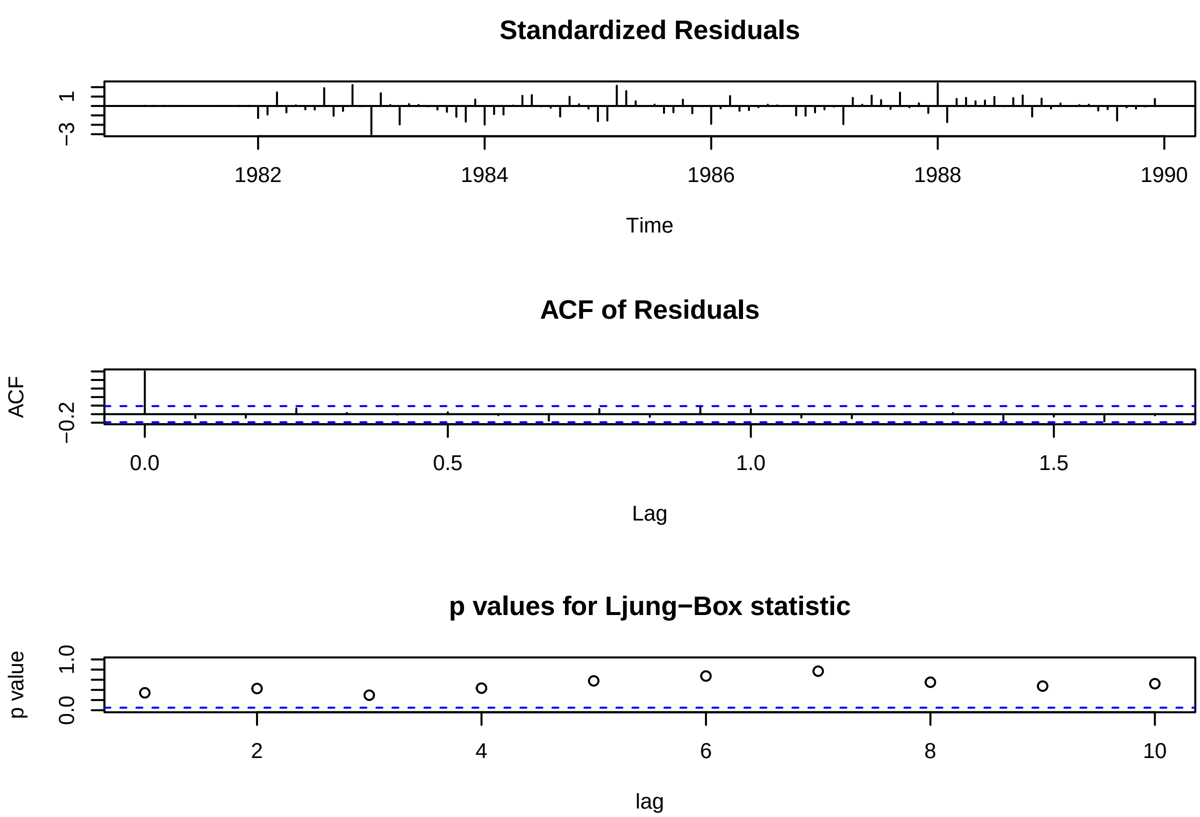
x

Call:
arima(x = training_ts, order = c(1, 0, 1), seasonal = list(order = c(0, 1, 1),
period = 12))

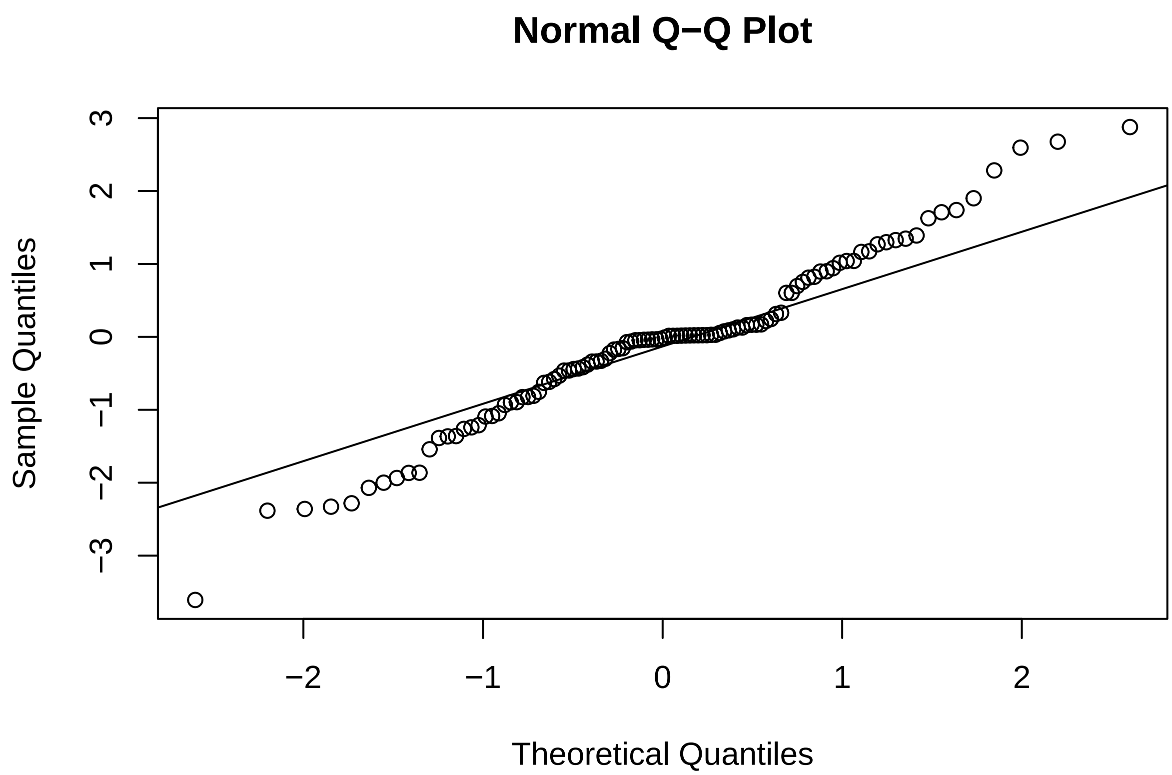
Coefficients:
ar1 ma1 sma1
0.9595 -0.8692 -0.9998
s.e. 0.0747 0.0898 0.2681

sigma^2 estimated as 1.443: log likelihood = -166.57, aic = 339.14

We then carry out the diagnosis of the residuals.



From the above plots, the distribution of standardised residuals appears to be random. The ACF plot also indicates that there is no significant correlations among the residuals, and the Ljung-Box statstics for different lags stay beyond the 0.05 significance line, further supporting the randomness of residuals.



Normality test of the residuals for this chosen model is also carried out through the QQ-plot and the Shapiro-Wilk test. The plot and a p_value of 0.04296 in the Shapiro-Wilk test suggest that the residuals of our chose model still deviates slightly from a normal distribution. The mean of the residuals (-0.088) is not very close to 0. We may just accept this to be a limitation of the model, since the small waving pattern identified in our earlier data exploration is not accounted for in this model.

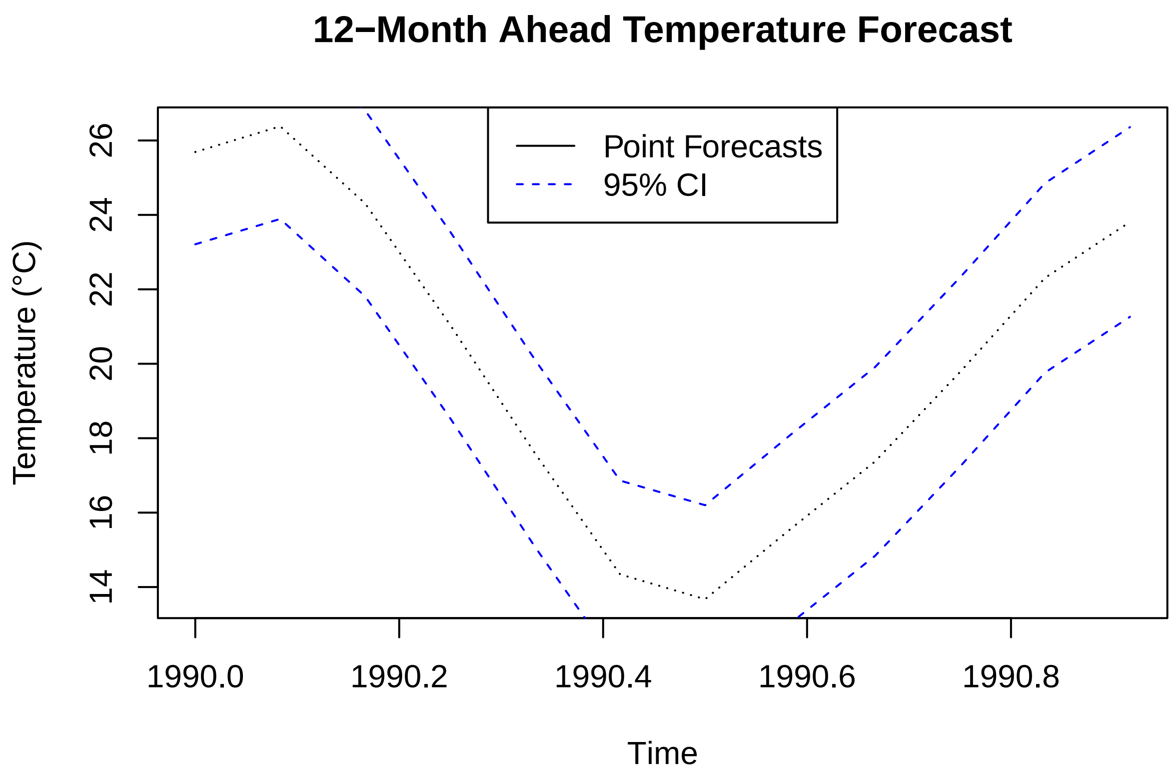
In the $SARIMA(1,0,1) \times (0,1,1)_{12}$ model adopted for the training set, there are:

- non-seasonal AR(1) and MA(1) items;
- a seasonal SMA(1) term and a seasonal difference term of $(1 - B^{12})$ (D = 1);

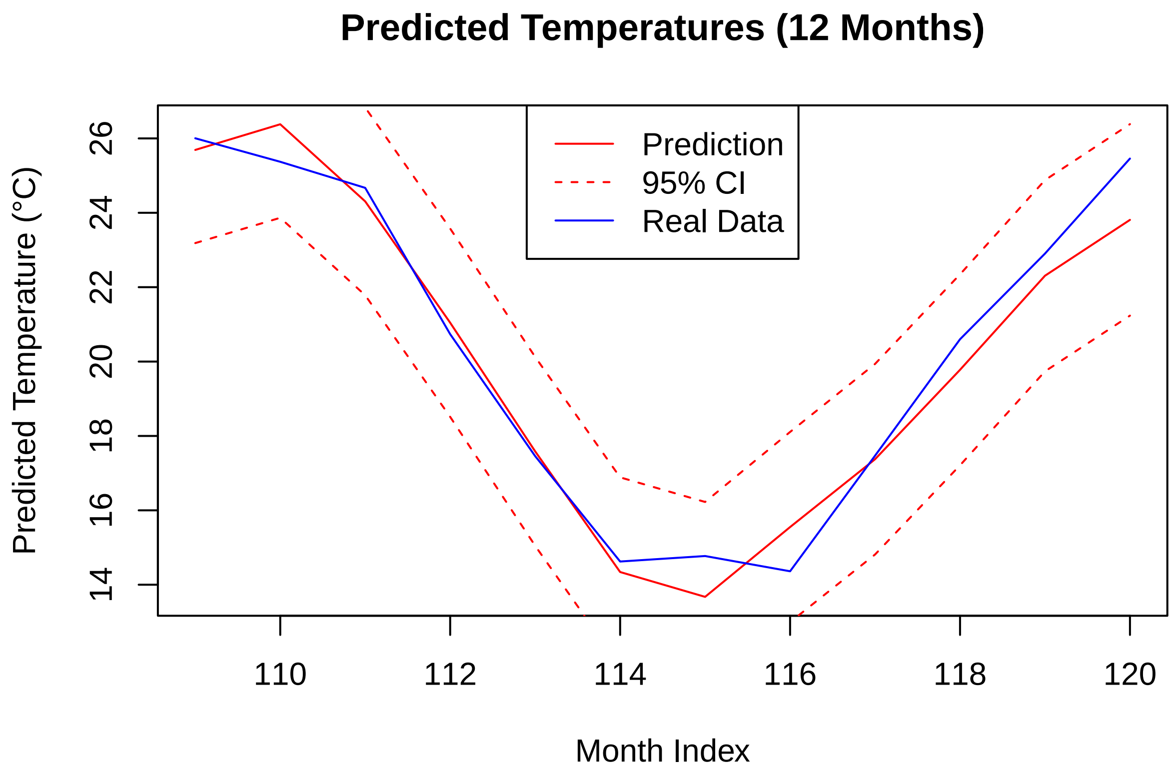
Therefore, the formula integrating the estimated coefficients for the model is $(1 - 0.9595B)(1 - B^{12})Y_t = (1 - 0.8692B)(1 - 0.9998B^{12})\epsilon_t$, which can be further expanded to find the formula for Y_t.

Forcast

Based on the chosen model, we now plot a forcast for the 12-steps ahead together with the lower and upper bounds of the 95% confidence interval.



To evaluate the predictions better, we zoom into the 12 month steps forcast, and check if the test data fits well between the confidence interval of the predictions.



As shown in the above graph, the real test data (the blue line) lies perfectly in the 95% confidence interval (two dotted red lines) of the model. Hence the chosen model works well for the test data, and this good performance also helps to justify our decision to treat the residuals of the model as roughly normal.

Conclusion

The exploration of the daily maximum temperatures in Melbourne clearly demonstrates a yearly cycle, and didn't identify any obvious long-term trend. In order to facilitate the analysis as well as reduce the noise, the daily temperatures were aggregated into monthly mean temperatures. Only a seasonal difference of the data is necessary for the monthly data. By comparing a few SARIMA models with different p, q and P, Q parameters, it has been concluded that a $SARIMA(1,0,1) \times (0,1,1)_{12}$ model works the best for the training set, and it also predicts well on the test set. There is a limitation to the chosen model though. The slight waving trend in the data was not taken into account, and the residuals of the model deviates slightly from a normal distribution. Incorporating a sine or cosine component into the model may help overcome this limitation and is worth further exploration.

References

Dahiya, P., Kumar, M., Manhas, S. et al. (2024). Time series study of climate variables utilising a seasonal ARIMA technique for the Indian states of Punjab and Haryana. Discover Applied Sciences, 6(650). https://doi.org/10.1007/s42452-024-06380-5