Temperature Trends in New Delhi

Jessica Bunge and Brandon Coates
MA5781 - Summer 2022

Abstract

Trends in temperature have long affected the running community, as marathon races are scheduled to optimize running conditions for elite runners. This study looks at the temperature trends in New Delhi, India, and seeks to determine if there is a linear deterministic trend in the temperature between years 2013 and 2017 along with using time series modeling techniques to forecast the ideal time of year to schedule New Delhi's marathon in the future. After determining that there is no deterministic linear trend in the data, an SARIMA$(1,0,0)(2,1,1)_{12}$ model was used to forecast the temperature both 10 and 20 years after data collection was ended in 2017. The findings indicate that January is and will continue to be the best month for marathon running, based on an ideal temperature of 10°C. While January is above this ideal temperature, it is consistently the coolest month in New Delhi based on seasonal trends and is therefore closest to the ideal temperature.

Introduction

Climate change affects many different aspects of life on this planet, and one of those aspects is the running community. Marathon runners in particular train months in preparation for their big days. Many marathon runners choose marathons specifically because they offer ideal running conditions. There are many factors that make conditions "ideal," and many are subjective to the individual. One particular factor is temperature. While the ideal temperature for marathon runners does have some variation, most running experts agree that it falls somewhere around 10°C for average runners. With temperature spikes and gradual changes over the years, it is possible that marathon majors will be changing dates within the next couple of decades to keep conditions prime for record times.

In this study, the city of New Delhi, India will be the focus. The study aims to answer two questions:

- Is there a linear trend in the average monthly temperature in New Delhi? If so, what is the rate of increase/decrease?
- Studies show that 10 degrees Celsius is the ideal temperature for many marathon runners. Based on this temperature, which month during our data window appears most optimal for marathon running? Does that optimal month remain consistent as we forecast 10 years in the future? 20 years?

These questions will be addressed using a variety of time series models and thoughtful model selection to forecast ten and twenty years out from the time the data was collected, ending in 2017. This report will outline the data itself, the tests and models considered, the selection process for the final model, and the forecasted future data.

Time Series Data

The data used from kaggle.com contained 1576 observations, one per day from January 1, 2013-April 24, 2017. Each observation included the mean temperature (in degrees Celsius) on that given day in New Delhi, India. From these daily observations, monthly means were calculated to put the data in the form to best answer the research questions. This yielded 51 observations of monthly averages. The final nine months were set aside for the test set, leaving the first 42 observations for the training set. This was used to check the selected model for overfitting or other potential concerns.

The data from the training set was used in fitting potential models. The time series plot is shown below in *Fig. 1.*

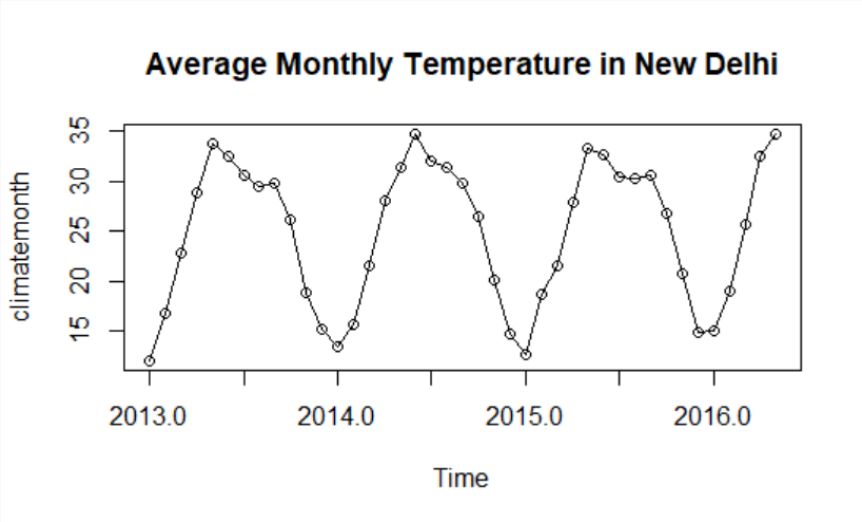

Fig. 1: Monthly Data Time Series Plot

Upon initial examination, the time series plot appears to show some clear seasonality along with a potentially deterministic upward trend over time.

Seeing the upward trend and seasonality, the next step was to use R to decompose the process, as seen in *Fig. 2*. This separated the random element, the seasonal element, and the polynomial trend for closer examination. Looking at the decomposition, the seasonal element matches the data closely, giving evidence that there is strong seasonality in the data. The polynomial trend appears to potentially follow an upward linear trend, though a quadratic element might give a better fit. The decomposition of this element does not necessarily show the trend to be deterministic, so further exploration was needed to decide if a linear or quadratic trend was needed.

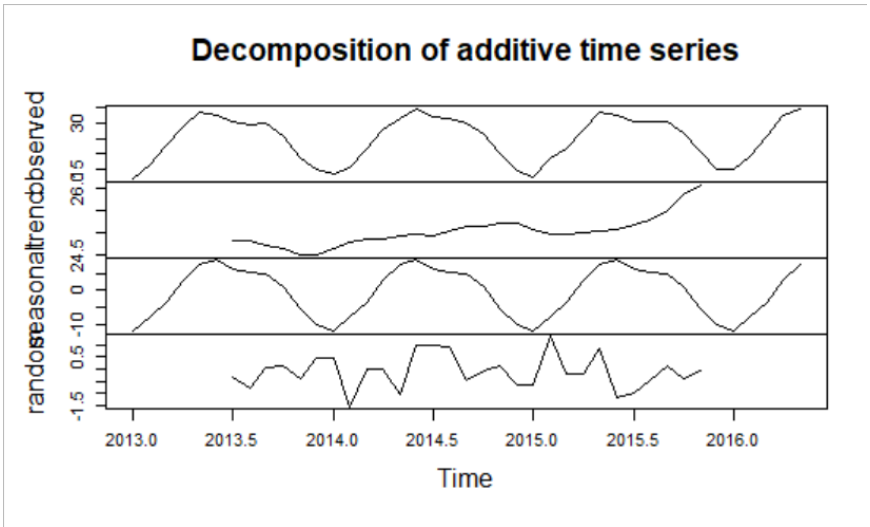

Fig. 2: Decomposition of Monthly Data Time Series

Potential Transformations

Log Transformation

Before beginning to explore models, transformations need to be considered. The first transformation to be examined was the log transformation. The main reason a log transformation is considered is to decrease changes in variance over time. Looking at the side-by-side plots in *Fig. 3*, the log transformation does not appear to make any significant changes in variance. Therefore, the log transformation was rejected for the sake of simplicity.

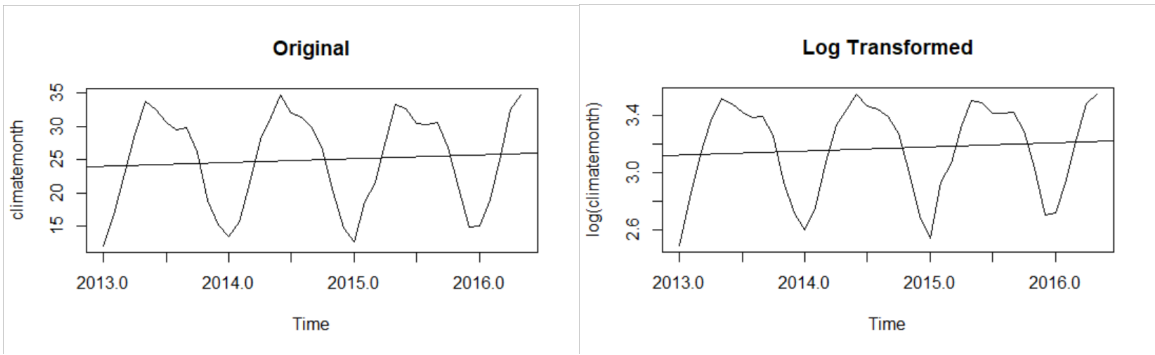

Fig. 3: Non-Transformed data vs. Log-Transformed Data

Difference

Because of the slight upward trend in our data that could imply a linear trend, the first difference of the data was also considered. The side-by-side plots in *Fig. 4* show this transformation. The differencing appears to give a slight downward trend. Due to the dangers of over-differencing, the original data was decided upon for model building with a linear trend built in to make differencing unnecessary.

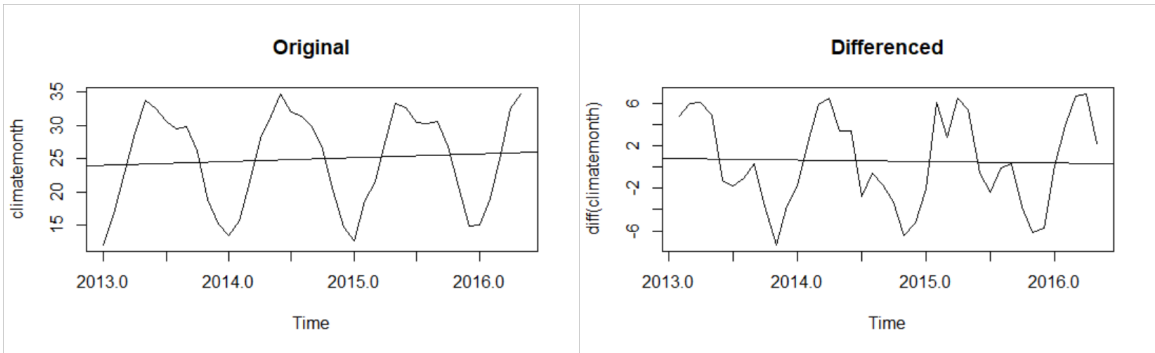

Fig. 4: Non-Transformed data vs. Differenced Data

Seasonal Means Model

The first model class considered, due to the clear seasonality in the data, was a seasonal means model. Seasonal means with a linear trend was compared to seasonal means with a quadratic trend, due to the indeterminate nature of the trend shown in the decomposition (*Fig. 2)*.

Linear Trend

The first model examined was a seasonal means model with a linear trend. The residuals (*Fig. 5)*, appear to be clustered around 0, though with some long runs on each side.

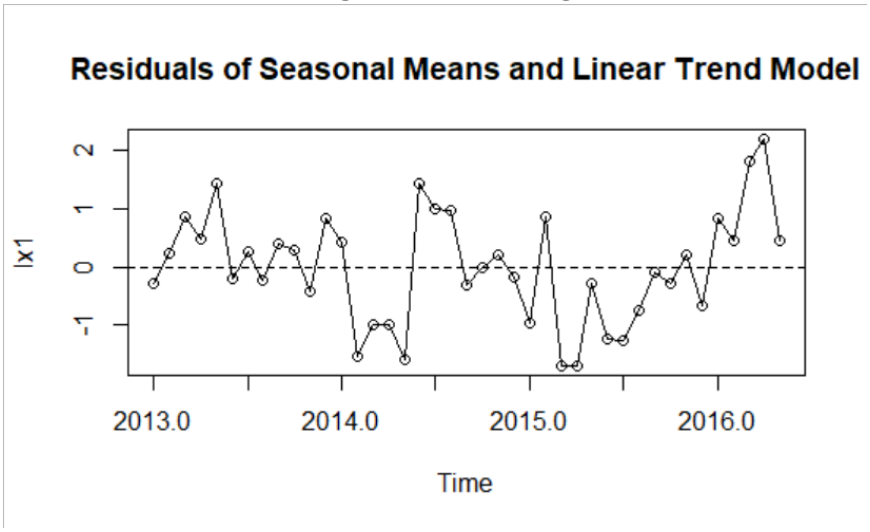

Fig. 5: Residuals of Seasonal Means with Linear Trend

The next step was to check for normality and independence of the residuals using the Q-Q plot (*Fig. 6*), Shapiro-Wilk test, and runs test (*Table 1*).

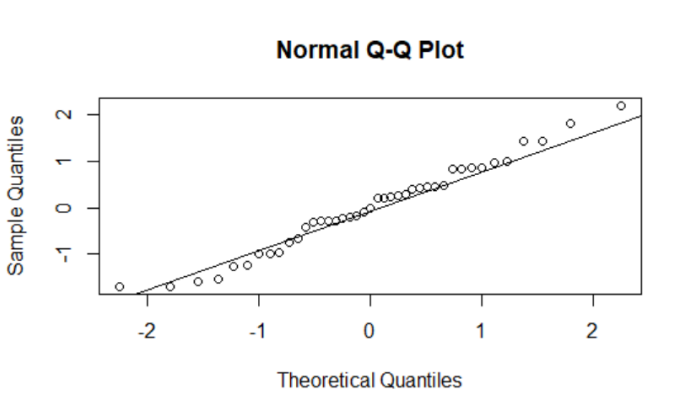

Fig. 6: Q-Q Plot Seasonal Means with Linear Trend

Test	H_0	H_a	p-value
Shapiro Wilk Test	Residuals are normally distributed	Residuals are not normally distributed	.6547
Runs Test	Residuals are random and independent	Residuals are not random or are dependent	.346

Table 1: Hypothesis Test Results - Seasonal Means with Linear Trend

The results of these tests show the residuals to be approximately normally distributed and independent.

Quadratic Trend

The next seasonal means model was built with an added second degree term to see if a quadratic trend was a better fit than a linear trend.

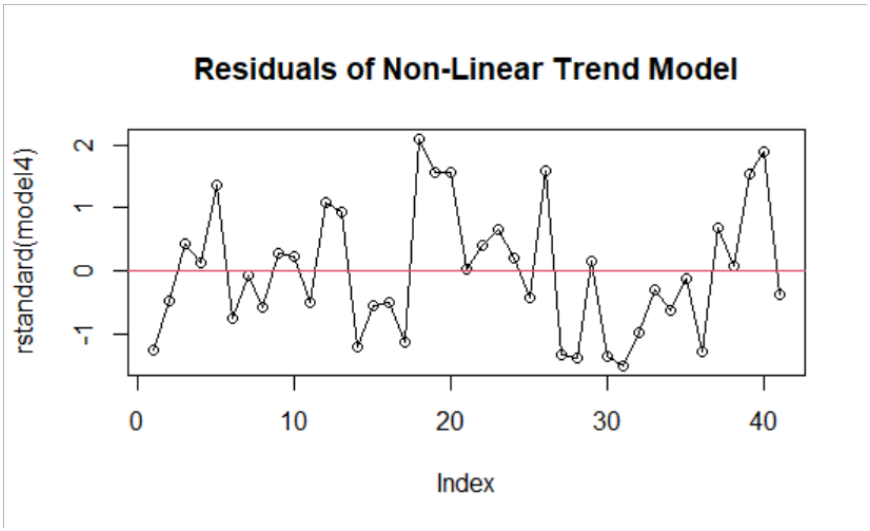

Fig. 7: Residuals of Seasonal Means with Quadratic Trend

Fig. 7 shows the residuals of this model to once again be centered close to 0 with no clear outliers.

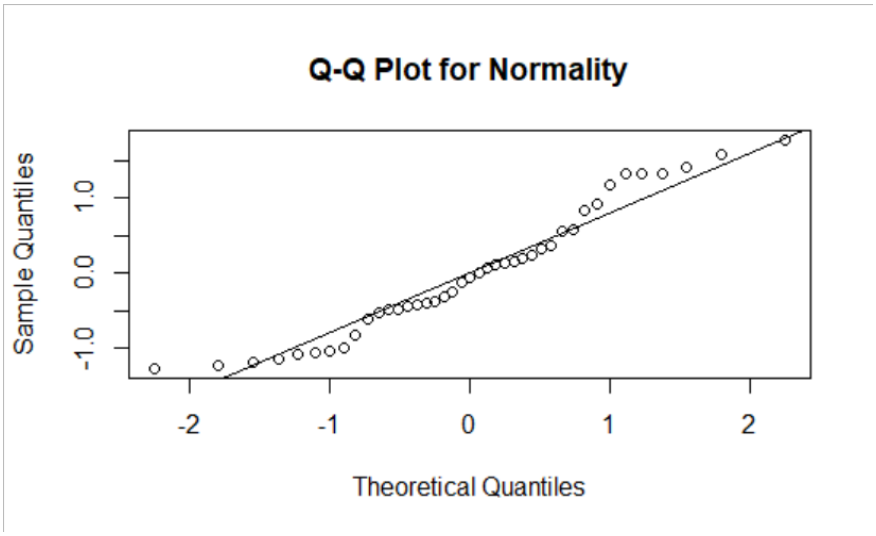

Fig. 8: Q-Q Plot Seasonal Means with Linear Trend

Test	H_0	H_a	p-value
Shapiro Wilk Test	Residuals are normally distributed	Residuals are not normally distributed	.07016
Runs Test	Residuals are random and independent	Residuals are not random or are dependent	.0562

Table 2: Hypothesis Test Results - Seasonal Means with Quadratic Trend

The analysis of residuals (*Fig. 8* and *Table 2*) show to be approximately normally distributed as well as independent.

However, the residuals from the seasonal means model with a linear trend look to be closer to normally distributed as well as more random, so if a seasonal means model is used throughout the rest of our model building, it will be paired with a linear trend.

Candidate Models

Deterministic Using Seasonal Means Model with Linear Trend

Using the residuals from the seasonal means model with a linear trend, ARMA models were considered. The plot of residuals (*Fig. 5*), as well as the ACF and PACF plots (*Fig. 9*) show significance in lag 1 and lag 12, after which the lag decays for each seasonal cycle. The EACF plot (*Fig. 10*) suggests an AR(1), MA(1), or an ARMA(1,1) model might be a good fit, so all three models were considered and compared.

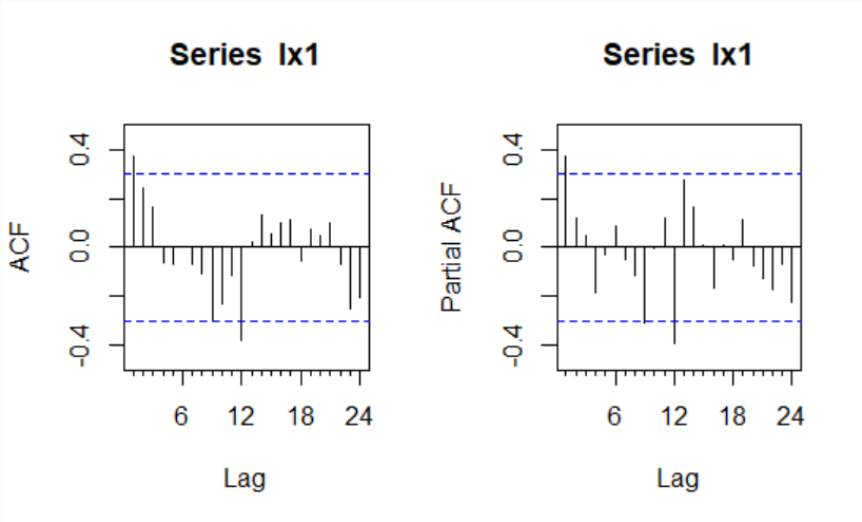

Fig. 9: ACF and PACF Plots - Residuals of Seasonal Means

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	o	o	o	o	o	o	o	o	o	o	x	o	o
1	o	o	o	o	o	o	o	o	o	o	o	x	o	o
2	x	o	o	o	o	o	o	o	o	o	o	o	o	o
3	o	x	o	o	o	o	o	o	o	o	o	x	o	o
4	o	x	o	o	o	o	o	o	o	o	o	o	o	o
5	x	o	x	o	o	o	o	o	o	o	o	o	o	o
6	x	o	o	o	o	o	o	o	o	o	o	o	o	o
7	x	o	o	o	o	o	o	o	o	o	o	o	o	o

Fig. 10: EACF Plot - Residuals of Seasonal Means

The augmented Dickey-Fuller test shows the data to be not stationary, with a p-value of 0.3546 while the Phillips-Perron unit root test shows enough evidence that the data might be stationary, with a p-value of less than 0.01. To avoid over-differencing, no *d* value was added into the models created.

Model	***Coefficient(s)***	***AIC***
MA(1)	θ_1= .3050	112.77
AR(1)	ϕ_1 = .3665	111.72
ARMA(1,1)	θ_1= -0.2562 ϕ_1 = .5877	113.51

Table 3: Comparison of ARMA Models on Seasonal Means Residuals

Looking for reasonable candidates in these three models, it appears that the AR(1) model has the lowest AIC value (*Table 3*), so AR(1) was kept as a candidate model for final model selection while the other two models were rejected.

Stochastic using Seasonal ARIMA Model with Period = 12

If the seasonality of the original data is considered to be stochastic instead of deterministic, a seasonal means model is not the best choice. Instead, a seasonal ARIMA model should be considered. Considering this, the next model built was an SARIMA model. Using the autoarima function in R, the best choice of p, d, and q gave an SARIMA(1,0,0)x(1,1,0)$_{12}$ model. Looking at the standardized residuals of this model (*Fig. 11)*, they appear to be randomly distributed around 0 after the first 12 month period. There are a few that fall outside of |2|, which could be a concern to address when determining the final model.

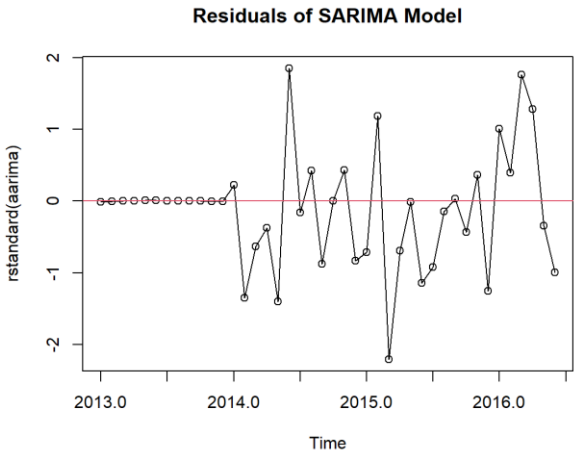

Fig. 11: Residuals of SARIMA Model

In addition to the shape of the residuals, tests for normality and independence were run.

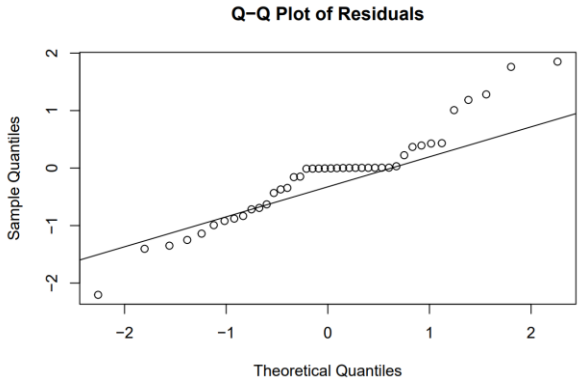

Fig. 12: Q-Q Plot SARIMA(1,0,0)x(1,1,0)$_{12}$

The Q-Q plot (*Fig. 12*) shows some slight skew on the left side but appears to be mostly linear.

Test	H_0	H_a	p-value
Shapiro Wilk Test	Residuals are normally distributed	Residuals are not normally distributed	.0476
Runs Test	Residuals are random and independent	Residuals are not random or are dependent	.465

Table 4: Hypothesis Test Results - SARIMA(1,0,0)x(1,1,0)$_{12}$

The Shapiro-Wilk test verifies the what is seen in the q-q plot, which is that there does appear to be a violation of the normal assumption, however, the sample size examined could be considered a large ($n > 30$) so this violation was not found to be an issue. The runs test indicates that it is safe to assume independence in the residuals. With all of this, the SARIMA(1,0,0)x(1,1,0)$_{12}$ on the residuals of a linear trend holds as a potential candidate model, along with the AR(1) model on the residuals of seasonal means.

Final Model Selection

Once the two candidate models were identified, over and under differencing was applied to the models to ensure that the optimum model was selected for this data set. This process consisted of altering the orders of the identified models by either increasing the order or decreasing the order of the model. For example, since one of the identified candidate models was an AR(1) model, an AR(2) model was examined to see if the information criteria values were less than the identified model. Several models were built to examine if the auto ARIMA function produced the optimum order of the SARIMA function. The orders were varied and the information criteria values were calculated for each model examined to examine if there was a more appropriate SARIMA model than the identified SARIMA(1,0,0)x(1,1,0)$_{12}$ model. A summary of several of the models examined is seen in *Table 5*¸ where it was observed that the SARIMA(1,0,0)x(1,1,1)$_{12}$ model resulted in low information criteria values than the identified candidate model. It should be noted that the SARIMA(1,0,0)x(1,1,1)$_{12}$ model did not have the lowest AIC values, the SARIMA(1,0,0)x(2,1,1)$_{12}$ model did, however this model added extra complexity which did not translate to a noticeable decrease in the AIC value, so in order to have a more parsimonious model with accurate prediction ability, the SARIMA(1,0,0)x(1,1,1)$_{12}$ model was selected as the final model.

Model	AIC	AICc	BIC
AR(1)	111.41	111.72	114.89
AR(2)	112.85	113.48	118.06
SARIMA(1,0,0)(1,0,0)$_{12}$	191.89	192.53	197.11
SARIMA(1,0,0)(1,1,0)$_{12}$	110.26	111.18	114.46
SARIMA(1,0,0)(1,1,1)$_{12}$	108.54	110.14	114.14
SARIMA(1,1,0)(1,1,1)	110.72	112.38	116.19
SARIMA(1,1,1)(1,1,0)$_{12}$	111.41	113.08	116.88
SARIMA(1,1,1)(1,1,1)$_{12}$	110.26	112.87	117.10
SARIMA(2,0,0)(1,1,1)$_{12}$	110.04	112.54	117.05
SARIMA(1,0,0)(2,1,1)$_{12}$	107.87	110.37	114.88

Table 5: Over and Under-Differencing Model Selection Results

Final Model Diagnostics

After selecting the SARIMA(1,0,0)x(1,1,1)$_{12}$ model as the final model, one last step must be taken before using the model to forecast on the data set. The model diagnostics must be examined to determine whether or not this model upholds the proper assumptions of normality, independence, and whether or not there is correlation between the model residuals that would lead to bias in the model. Diagnostics plots were created to examine if these assumptions held true for the final SARIMA(1,0,0)x(1,1,1)$_{12}$ model, which can be seen in *Figure 13*. When assessing the plot of the residuals, it is found that the standardized residuals were randomly scattered around zero and all of the standardized residuals fell within the range of -3 to 3, thus the normality assumption was found to be upheld for this model. The ACF plot that was created shows no points in which the ACF value of the residuals would be considered significantly different than zero outside of the lag of 0, so it was concluded that the assumption of independence holds true for this model. Lastly, a plot of the p-values for the Ljung-Box test were plotted to examine the autocorrelation between the model residuals. The Ljung-Box test tests the null hypothesis that the residuals are uncorrelated against the alternative hypothesis that the residuals are not uncorrelated by examining the autocorrelation functions of the residuals. In the Ljung-Box test plot, it can be seen that there are no p-values that were calculated that have a value of less than 0.05, which would reject the null hypothesis, thus, it was concluded that the residuals were uncorrelated, so the model was a proper model for forecasting this data set.

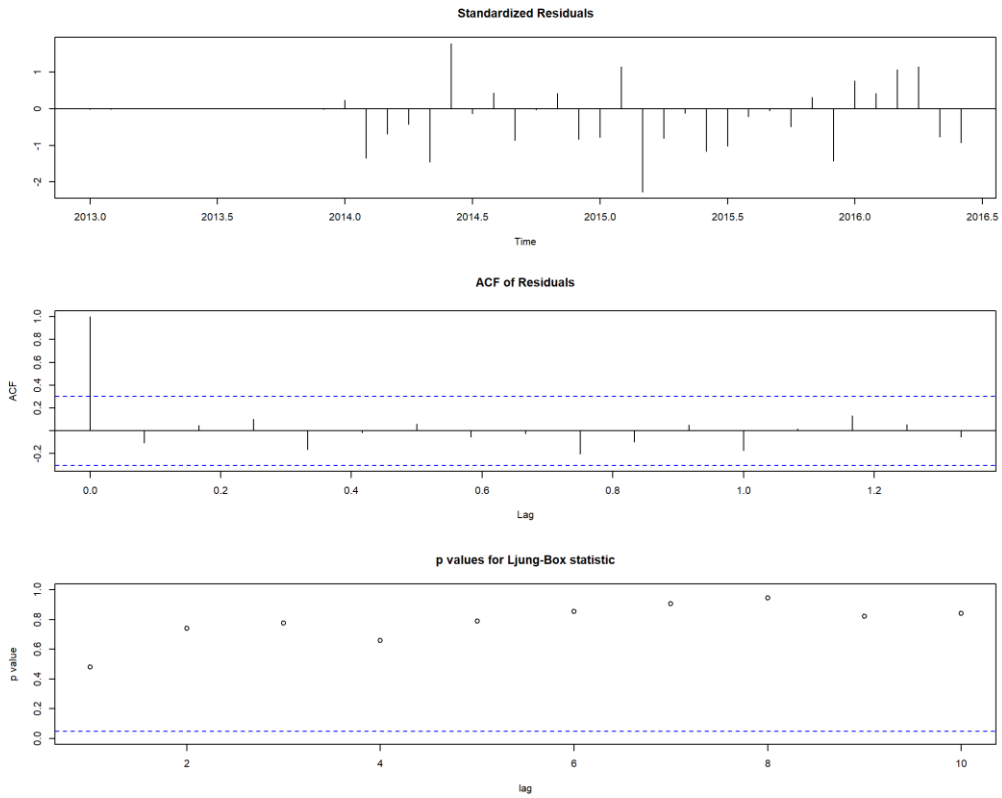

Fig. 13:Diagnostics Plots for SARIMA(1,0,0)x(1,1,1)$_{12}$

Forecasting

In order to determine the accuracy of the SARIMA(1,0,0)x(1,1,1)$_{12}$ model that was chosen as the final model, the model was used to forecast the data that was included in the test set that was held out of the training data set. The forecasted values can be seen in the time plot in *Figure 14*, which shows the actual recorded values in red, the 95% confidence interval bounds in green, the forecasted values in light blue, and the dashed line represents the optimum running temperature. It was found that the model was accurately able to forecast on the test set, with a root mean square error of 1.166 and mean absolute percentage error of 3.902%. This can be observed since the red and blue trendlines representing the actual data and forecasted data, respectively, follow each other closely with any deviations being captured within the 95% confidence interval bounds.

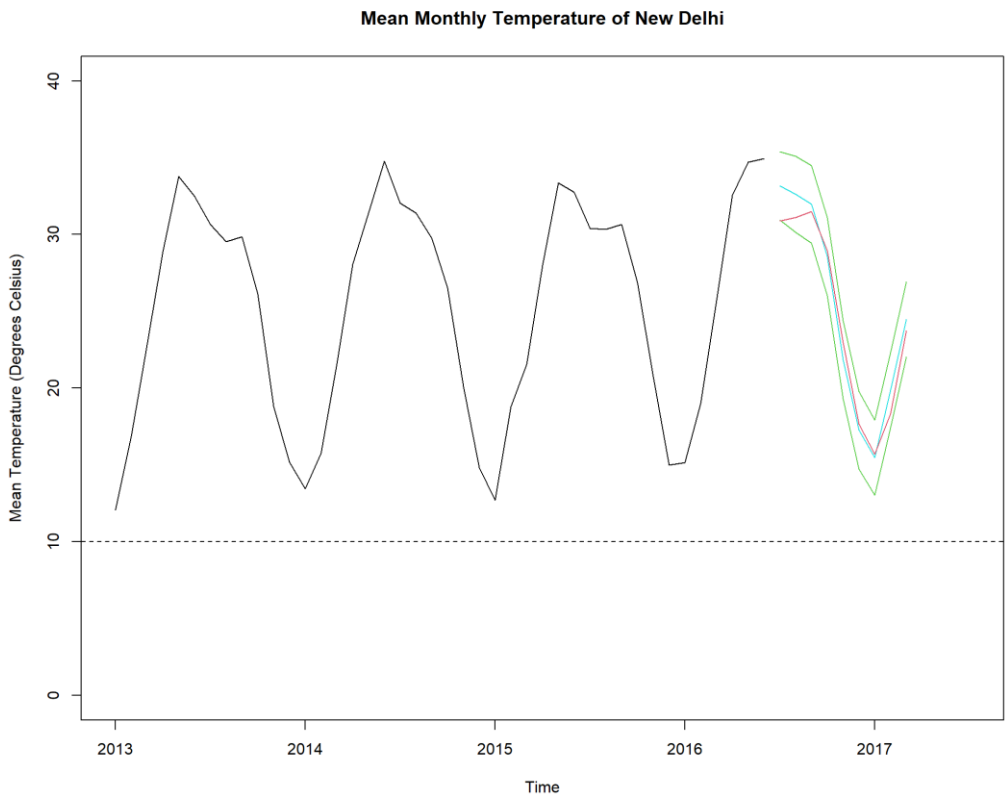

Figure 14: Time Plot of Forecasted Results compared to Recorded Values

After assessing how accurate the final model performed on the test data set, the model was used to predict the intervals of 10 years out from the end of the reported data and 20 years out from the end of the reported data. The forecast of the mean monthly temperature of New Delhi for 10 years after the data was recorded can be seen in *Figure 15*, in which the red trendline is the forecasted mean monthly temperature and the green trendlines are the 95% confidence interval bounds. It was observed that the forecasted values tend to follow the overall seasonal trend that was observed in the original data, however, the values every year were found to be slightly increasing as the time increased.

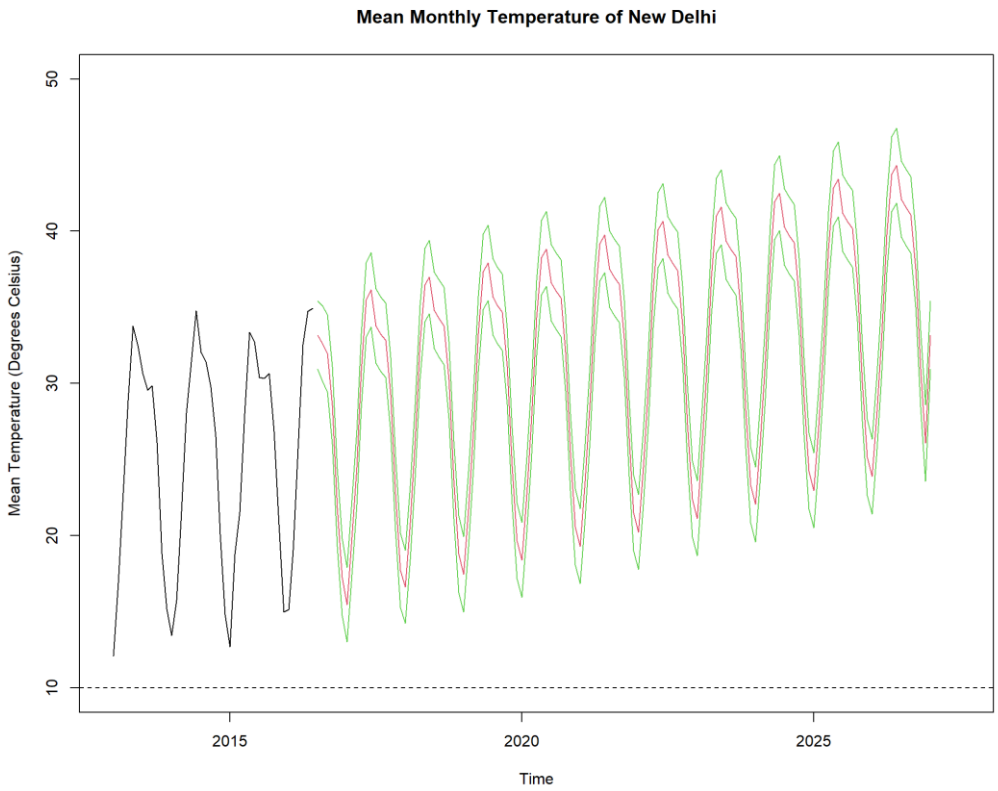

Figure 15: Time Plot of Mean Monthly Temperature of New Delhi Forecasted for 10 Years

The forecasted values for the next 20 years after the data was recorded follows a similar trend as was seen in the forecasted values for the next 10 years. The time series plot including the values of the mean temperatures that were forecasted for 20 years out can be seen in *Figure 16*. Once again, in this plot the red trendline is the forecasted value and the green trendlines are the 95% confidence interval bounds.

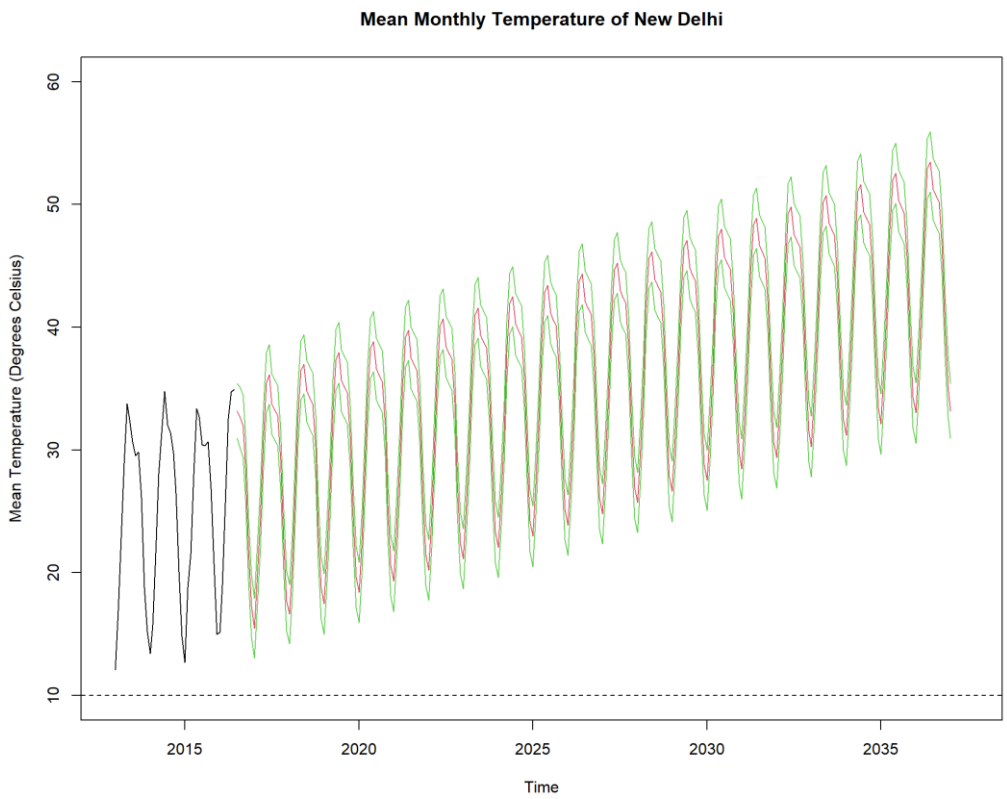

Figure 16: Time Plot of Mean Monthly Temperature of New Delhi Forecasted for 20 Years

From these plots, it was observed that while there is an increase in temperature year after year in terms of mean monthly temperature, it is not necessarily a linear trend. There is a significant seasonal component factor that must be considered for the mean monthly temperatures and it was found the optimum model, given our training data, factors in autoregressive, seasonal autoregressive, and seasonal moving average behaviors. Thus a piecewise linear trend may be able to be used to approximate the mean monthly temperature change from one year to the next, but a general linear trend may lead to erroneous forecasting results. It was also observed that in both the 10 year forecast and 20 year forecast, January was the best month to run a marathon in terms of only mean monthly temperature. It was found that January was consistently the month with the lowest temperature every year, which made it the month that was closest to the optimum running temperature of 10°C.

Limitations

Though the model was found to accurately forecast the mean monthly temperatures on the test set that was held out of the training data, there may be larger trends that may not have been included in the model that would lead to erroneous forecasts over long periods of time. Since the model was only trained on 42 months, there may be larger trends that were not included in the data which would be common in other climate change models that would examine a much larger amount of average monthly temperatures since climate change historically occurs over large periods of time. However, for short forecasting periods this model may provide sufficient forecasts since it did in fact still capture a seasonal trend that was increasing in general year

after year. This possible issue could be examined further in a follow-on effort if more data was gathered to test the model forecasting accuracy. It should also be noted that while the final model examines only the mean monthly temperature, it may not accurately reflect the best month for running a marathon in New Delhi since other factors such as air quality, wind, and humidity all can impact a runner's performance. A more complex multivariate model should be considered to factor in these variables to determine if January is truly the best month to hold a marathon in New Delhi. There may in fact be other months in which the mean temperature may be slightly hotter, but the other factors may lead to better running conditions.

Conclusion

The mean monthly temperature of New Delhi, India was examined for a time period of January, 2013 to April, 2017 to determine if time series analysis techniques could be utilized to identify the trend in temperature changes in this time period as well as identify the month that exhibits optimum running temperatures for holding a marathon in New Delhi in the given time period, over the next 10 years, and the next 20 years. The goal was to examine how climate change was effecting the mean monthly temperatures and if there were large changes that would shift the seasonality of average monthly temperatures year after year. It was found that a $SARIMA(1,0,0)x(1,1,1)_{12}$ model was able to accurately forecast the average monthly temperatures of a test set of 9 months that was held out of the original data set with an RMSE of 1.181. This model was used to forecast the average monthly temperatures in New Delhi to January 2027 and to January 2037. In these forecasted results, it was found that the increase average monthly temperatures we not necessarily linear, instead, it was found that a linear component was used, however, seasonality played a large role in determining the average monthly temperatures and that moving average and autoregression factors needed to be considered to accurately forecast the changes in average monthly temperature. A linear trend may give a rough approximation as to what the temperature may be, however, for more accurate forecasting, the more complex factors would need to be examined. It was also found that when examining only average monthly temperatures in New Delhi, January consistently appeared to be the month the exhibited temperatures as close to the optimum running temperature of 10°C. This was not only true in the data that was used to train the model, but true over the forecasted values up to January, 2027, as well as the forecasted values up to January, 2037. However, it should be noted that this model only examines the mean monthly temperature of New Delhi, India and not other factors that may impact a runner's performance such as air quality, wind, and humidity.

References

<https://www.kaggle.com/datasets/sumanthvrao/daily-climate-time-series-data>
Data used under CC0 license in Public Domain