MA 585 Final Project

Hsueh-Pin Liu

2023-04-28

Introduction

Climate change has become a major concern globally, and Brazil is one of the countries that is most vulnerable to its impacts. As a result, it is important to understand how climate change is affecting different regions of Brazil. This report will focus on the city of Rio de Janeiro and its temperature trends over the past few decades. Temperature is a critical indicator of climate change, and studying its patterns in Rio de Janeiro can help us better understand the impacts of climate change on urban areas in Brazil. Rio de Janeiro is a major city located on the coast of Brazil, with a tropical savanna climate. Due to its location, it is susceptible to rising sea levels and extreme weather events caused by climate change. This report will examine the temperature trends in Rio de Janeiro over the past few decades, identify any significant changes in temperature, and explore the possible drivers of these changes. By studying the temperature patterns in Rio de Janeiro, we can gain insights into the broader impacts of climate change in Brazil and inform policies and adaptation efforts to mitigate its negative effects.

Dataset

The Brazil Weather Information dataset by INMET, available on Kaggle, provides a comprehensive source of weather data from various locations in Brazil. This report will use the data from the weather_sum_2020 to weather_sum_2023 files, which provide summary statistics for each day of the year, to analyze the temperature patterns in Brazil during the summer months. I use the data from 2020-2022 as train data and the 3 months data from 2023 as test data. By examining this data, we can gain insights into the impacts of climate change on Brazil's temperature and other weather variables. After taking steps to clean the data(Appendix 1), here are the first six rows of the data.

```
## # A tibble: 6 x 11
##   ESTACAO DATA (YYYY-M~1 rain_~2 rad_max temp_~3 temp_~4 temp_~5 hum_max hum_min
##   <chr>   <date>         <chr>   <chr>     <dbl> <chr>   <chr>   <chr>   <chr>
## 1 A636    2020-01-01     0       3728.6     28.2 35.7    21.3    90      32
## 2 A636    2020-01-02     0.8     1651.6     26.1 30.9    22.7    89      48
## 3 A636    2020-01-03     12.8    1879.2     23.5 28.1    21.2    91      62
## 4 A636    2020-01-04     0.4     2633.3     24.5 28.9    21.1    91      58
## 5 A636    2020-01-05     1.4     2715.3     24.8 28.9    22      91      64
## 6 A636    2020-01-06     13      3712.1     26.1 33.6    21.7    89      46
## # ... with 2 more variables: wind_max <chr>, wind_avg <chr>, and abbreviated
## #   variable names 1: `DATA (YYYY-MM-DD)`, 2: rain_max, 3: temp_avg,
## #   4: temp_max, 5: temp_min
```

Basic visualization of the data

To gain a better understanding of the temperature trends, I created two time series plots, one using the average temperature and the other showing the highest and lowest temperatures (see Fig 1). These plots appeared stationary and indicated obvious seasonality. To further analyze the data, I used a decomposition plot and a Dickey-Fuller test (refer to Appendix 2). The decomposition plot revealed a clear trend and seasonal pattern, with a seasonal period of 365 days due to the daily nature of the data. Additionally, the Dickey-Fuller test produced a p-value of 0.01, leading to the rejection of the null hypothesis and indicating that the data is indeed stationary. Therefore, no further transformation of the data is required.

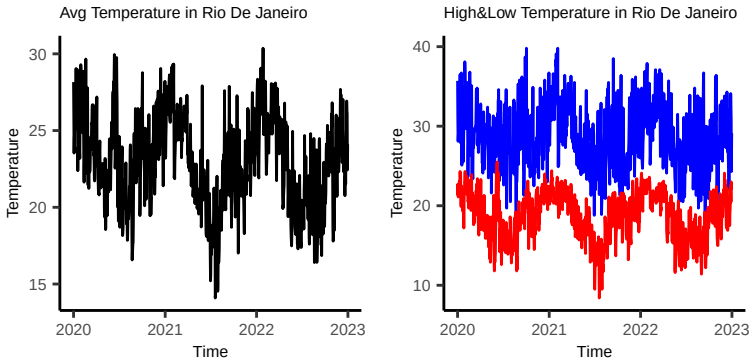


Figure 1: Average temperature(Left)&Highest and Lowest temperature(Right) Time Series Plot

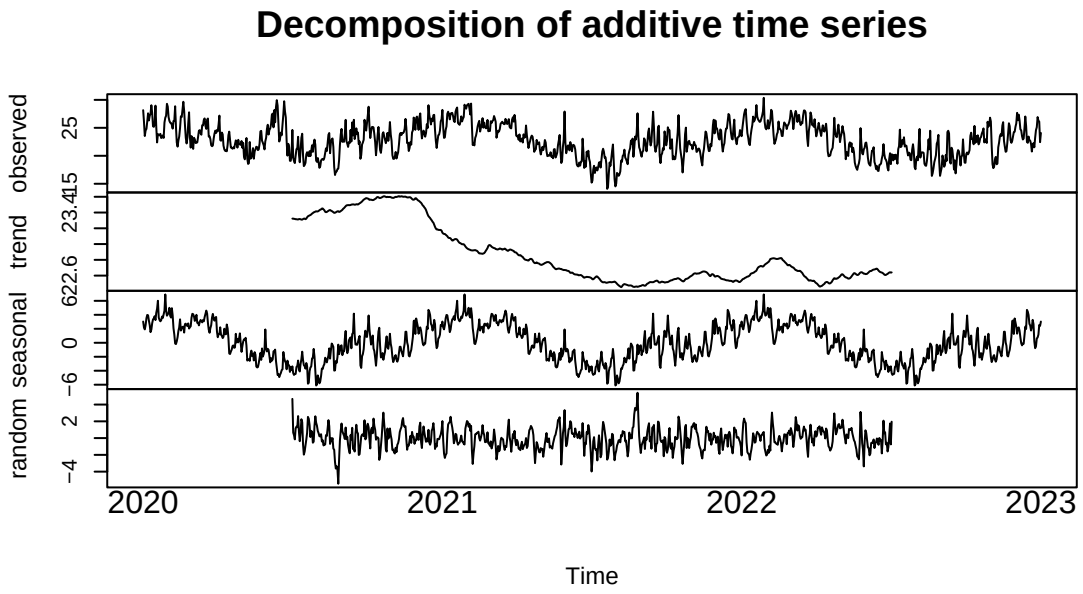


Figure 2: Decomposition Plot

Modeling

Model-based Forecast

Choosing the appropriate model

To forecast the time series data, I analyzed its ACF and PACF plots. The highly autocorrelated nature of the data made it challenging, so I tried different methods like log transformation(Appendix 3) and differencing(Appendix 4). However, none of them proved to be effective, so I decided to build the model using the original data despite its autocorrelation(Fig 3).

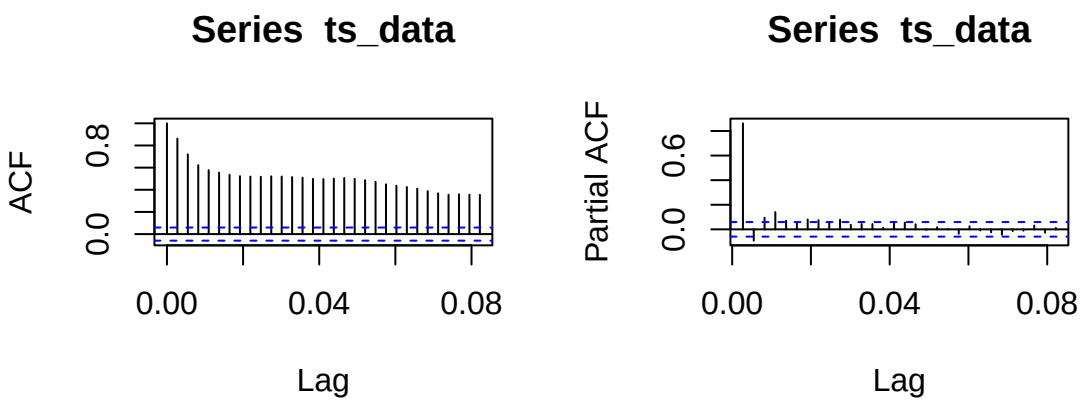


Figure 3: ACF and PACF Plots

Due to the seasonality present in the data, I have employed an SARIMA model to capture its patterns. To identify the optimal model fit, I have utilized the auto.arima function, and the resulting output suggests that the most suitable model is SARIMA $(2,0,0) \times (0,1,0)_{365}$, with the lowest AICc value. Hence, I will be using this model to forecast the time series.

Model	AICc
$(1,0,0) \times (0,1,0)_{365}$	2280.888
$(2,0,0) \times (0,1,0)_{365}$	2622.169
$(3,0,0) \times (0,1,0)_{365}$	2265.05
$(2,0,1) \times (0,1,0)_{365}$	2264.194
$(1,0,1) \times (0,1,0)_{365}$	2263.032
$(3,0,1) \times (0,1,0)_{365}$	2265.535

Forecasting

And then I use the model to forecast the temperature in the next three months(Fig 4) in order to compare with the test data, and the result seems reliable.

Diagnostics

The diagnostic plots(Fig 5) indicate that the residuals and p-values are not significant, suggesting that the model is appropriate.

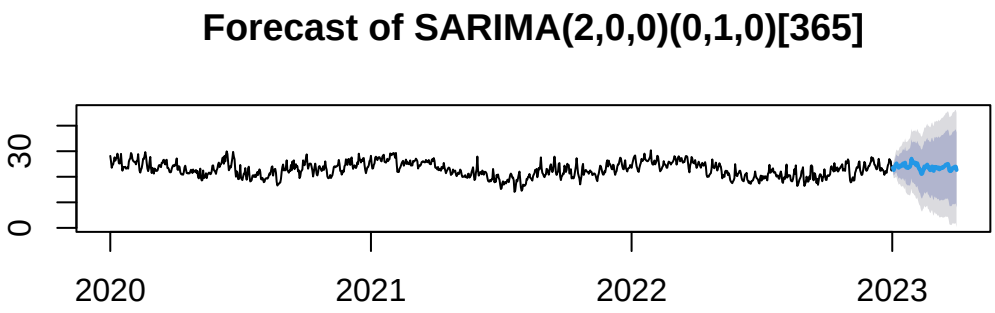


Figure 4: Forecast of the Model

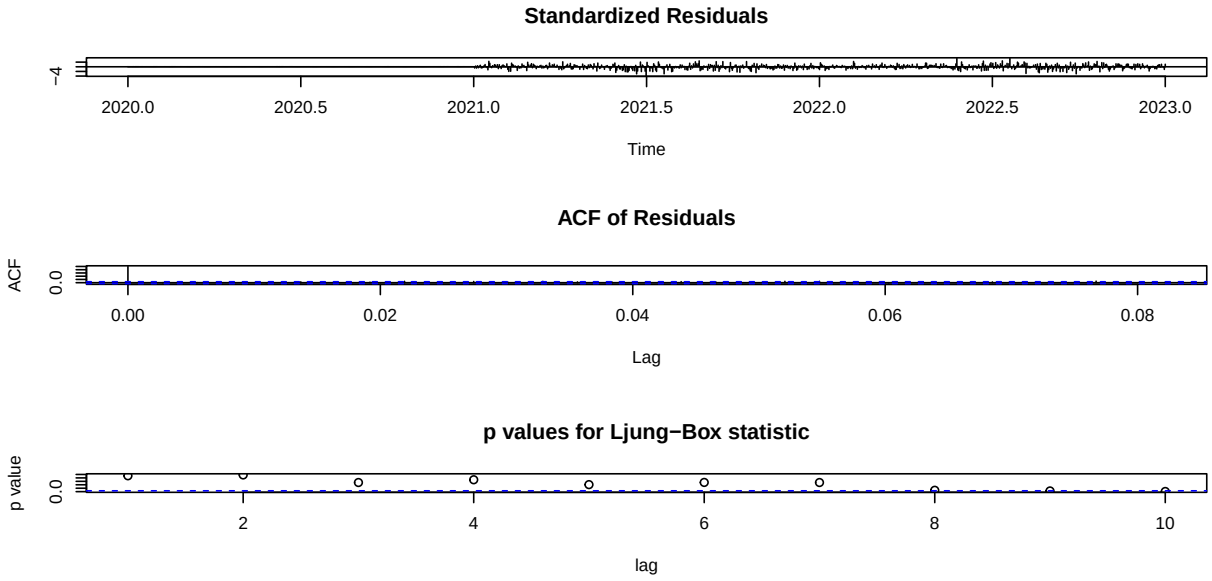


Figure 5: Diagnostic Plots

Holt-Winters Forecasting

Given the clear periodicity of the data, utilizing the Holt-Winters smoothing technique to forecast the data is a dependable approach. By incorporating the average temperature, a smoothing model can be generated to accurately predict the temperature. (Fig 6) The plot also looks reliable.

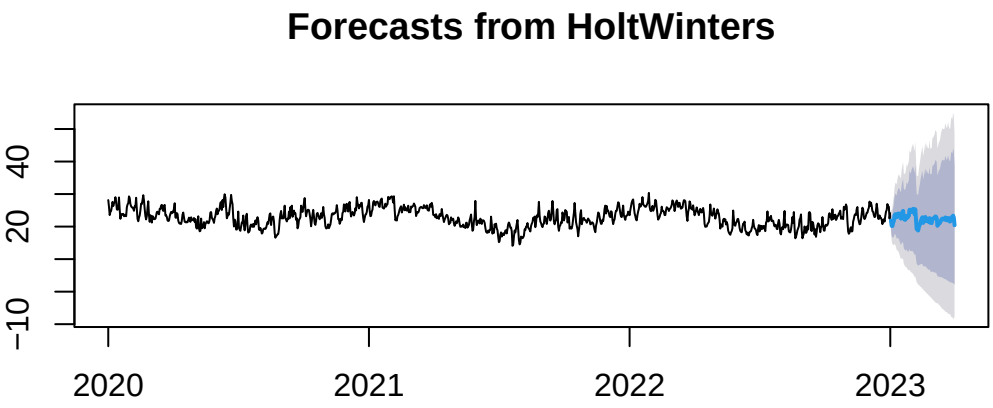


Figure 6: Holt-Winters Forecasts

Compare the forecasts accuracy

Both the SARIMA and seasonal Holt-Winters models appear to be viable options for forecasting. However, to determine which model performs better, we need to evaluate their MAE, RMSE, and MAPE. After analyzing the data, we can see from the table below that the SARIMA model outperforms the Holt-Winters model in terms of accuracy.

Criteria	SARIMA	Holt-Winters
MAE (Mean Absolute Error)	2.352121	3.556123
RMSE (Root Mean Squared Error)	2.721319	2.973317
MAPE (Mean Average Percent Error)	9.026348	13.590407

Conclusion

Based on the analysis of the average temperature in Rio de Janeiro from 2020 to 2023, it was found that the SARIMA model performed better than the Holt-Winters method in forecasting future temperatures. The comparison of MAE, RMSE, and MAPE for both methods supported this conclusion. For future analysis, it is recommended to focus on the highest and lowest temperatures to gain a better understanding of the climate in Rio de Janeiro. This can be done by analyzing extreme events and their impacts on the local environment, such as heat waves and cold spells.

References

1.https://www.kaggle.com/datasets/gregoryoliveira/brazil-weather-information-by-inmet?select=stations.csv

2.https://blog.csdn.net/cl1143015961/article/details/44982691

3.https://www.cnblogs.com/statruidong/p/6902315.html

Appendix

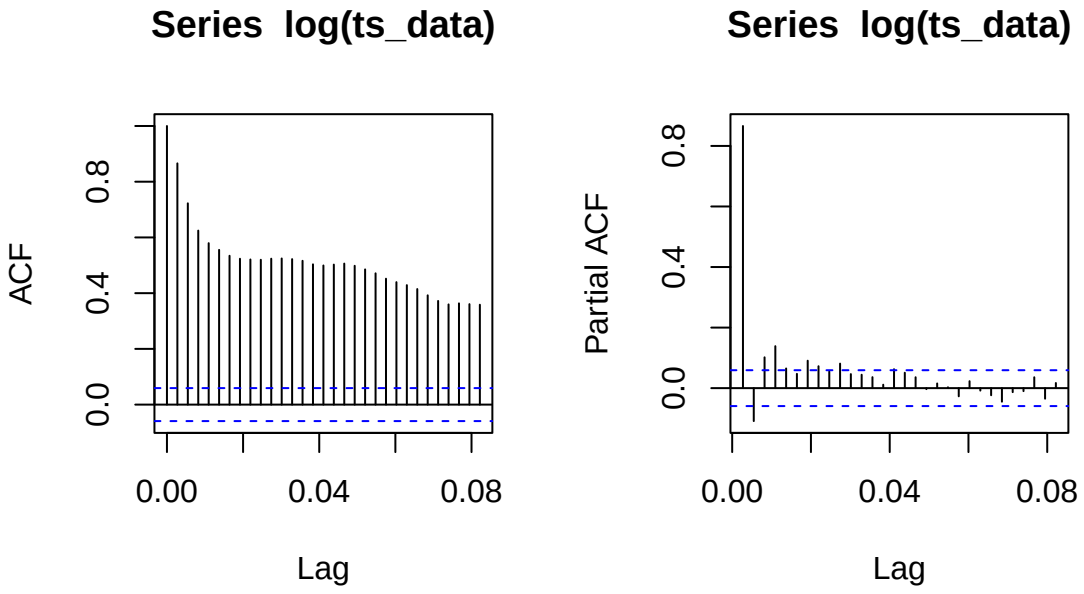
1

First I combine the data from 2020 to 2022 together, then I think it's strange that there are only 1095 observations while there are 1096 days, so there must be some data missing, after looking into the data, I don't have the data of 2020-06-14, so I use the average data of 2020-06-13 and 2020-06-15 as the data of 2020-06-14 of A636. Then I use the data from 2023 as the test data.

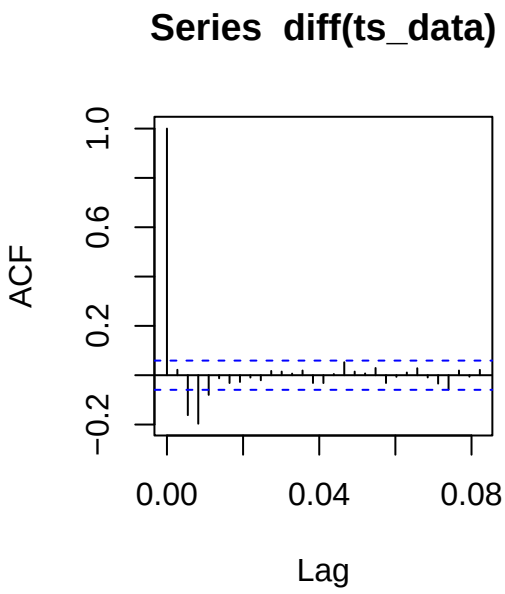
2

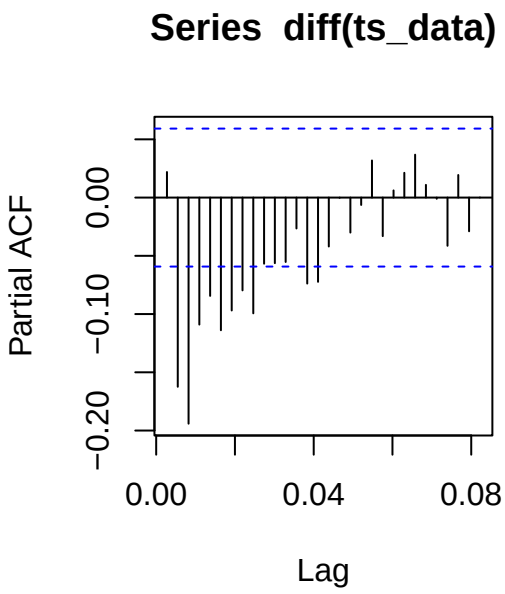
```
##
##  Augmented Dickey-Fuller Test
##
## data:  ts_data
## Dickey-Fuller = -4.1306, Lag order = 10, p-value = 0.01
## alternative hypothesis: stationary
```

3



4


Series diff(ts_data)
ACF
1.0
0.6
0.2
−0.2
0.00
0.04
0.08
Lag


Series diff(ts_data)
Partial ACF
0.00
−0.10
−0.20
0.00
0.04
0.08
Lag