SOUTH DAKOTA
STATE UNIVERSITY

Time Series Analysis: Final project

Group Members: Schuylar Harris Donovan, Michael Abalo, Theophilus Anim Bediako

December 11, 2022

Exploring Different Time Series Models

In this project, we explore different time series models using data from the National Oceanic and Atmospheric Administration (NOAA). The data shows records on snow depth, precipitation, snowfall etc from one of the testing locations in South Dakota from 1990-01-01 to 2012-12-31. We chose the maximum temperature variable to perform a univariate time series analysis since that had a few missing records as compared to the other five numeric variables. Out of the 8035 days within this period, the data only has 7763 records. We examined the missing records and found the missingness to be due to randomness(refer to Table 1 and Table 2). To impute the missing records, we relied on the Missing Value Imputation by Kalman Smoothing and State Space Models (na_kalman) function within the imputeTS library. This is an algorithm for estimating a missing time series value based on available data up to time, t via the Kalman filter [1] and [2]. The Kalman filter relies on a series of mathematical models to estimate the next missing series as $x_{t+1} = Ta_t + K_t(y_t - Za_t)$. An extensive interpretation of this model can be seen in [2] and [3].

We begin the analysis by looking at the time series plot of the maximum temperature series(has 8035 records now). From Figure 1, the series reveals an erratic pattern throughout the years, a behavior that is common to temperature series. The erratic pattern is due to the seasonal effect in the series. The time series plot does not reveal any increasing trend. The results can be seen in the decomposition plot. This clearly suggests an additive model of the form in equation (1) may be appropriate for the series, where s_t is the seasonal effect, m_t is the trend and z_t is the error term [4].

$$x_t = m_t + s_t + z_t \tag{1}$$

The first model we considered is the harmonic model. We fitted an initial harmonic model with 6 cycles to the series. A harmonic model is able to account for smooth variation in the seasonal effects. A t-ratio at an approximate 5% significant level was used to identify terms with significant coefficients. We defined significant coefficients as having magnitudes of the t-ratio to be at least 3. Eight terms were identified to have significant coefficients. We fitted a new model with only the significant terms. The new harmonic model had a lower Akaike information criterion(AIC) compared to the initial harmonic model. Figure 2 shows the correlogram plot of the residuals of the new harmonic model. Clearly, the residuals are not white noise. Hence, the appropriateness of the model is questionable.

The next model we considered is the auto regressive moving average (ARMA) model. The best-fitting ARMA(p, q) is chosen using the smallest AIC by trying a range of combinations of p and q in the arima function with the help of a for loop. The best ARMA model was identified to have $p, q = 2$. The correlogram of the residuals of the ARMA$(2, 0, 2)$ model were identified to be a realization of white noise. The correlogram of the residuals squared showed similar results.

We further explored the autoregressive integrated moving average ARIMA. The ARIMA (p, d, q) process was fitted to the maximum temperature series using the arima function in R. Several p, d, q values with an upper bound of 1 were tried with the help of a for loop. We identified an ARIMA$(1, 1, 1)$ model to be the one with the lowest AIC. The correlogram of the residuals of the best-fitting model does not differ from white noise. The correlogram of the squared residuals also appear to be a realization of white noise.

The fourth model we considered is the seasonal ARIMA model or the SARIMA model. This model is an extension of the ARIMA model. The model incorporates the seasonal components that is not considered in the regular ARIMA model. According to [4], the seasonal $(p, d, q)(P, D, Q)_s$ can be expressed using the backward shift operator as seen in equation (2)

$$\Theta_P(\mathbf{B}^s)\theta_p(\mathbf{B})(1-\mathbf{B}^s)^D(1-\mathbf{B})^d x_t = \Phi_Q(\mathbf{B}^s)\phi_q(\mathbf{B})w_t \tag{2}$$

where $\Theta_P, \theta_p, \Phi_Q, \phi_q$ are polynomials of orders P, p, Q and q respectively. We specified the conditional sum-of-squares ("CSS") as our fitting method and fitted a range of SARIMA models. The best-fitting SARIMA model selected was the one with the lowest AIC. The order returned from the best fitting model was (0,0,0)(2,0,2). The correlogram of the residuals of the model appear to be a realization of white noise(see Figure 4).

The last model we considered is the GARCH model. This model is a generalized version of the autoregressive conditional heteroskedastic (ARCH) model. Usually, GARCH models are able to account for volatility in the variance.Based off of the correlograms, the residuals of the GARCH model did not appear to be like white noise. This leads us to believe that this model might not be the best fit.

Conclusion

To decide on the choice of the best-fitting model, we considered both the correlogram plots and the AICs of the selected models. With the exception of the harmonic and the Garch models, the other models had correlogram plots depicting white noise. The ARMA model produced the least AIC (58627.91) as seen in Table 3. We, therefore, settled on this as our best-fitting model. Our choice of the best model is solely based on correlogram plots and AIC. More extensive approach and model exploration may be required to determine appropriateness of the ARMA model for the maximum temperature series. The final ARMA model is represented in (3), where, w_t is the white noise.

$$x_t = 1.4316x_{t-1} - 0.4360x_{t-2} - 0.6715w_{t-1} - 0.1790w_{t-2} \tag{3}$$

In the R markdown file, we made prediction for maximum temperature in 2022 using the best model selected - ARMA(2,0,2).

Table 1: Number Missing days by years

Year	Number of Days
2001	30
2008	61
2009	30
2010	31
2011	120

Table 2: Missing days by Months

Month	Number of Days
April	30
February	28
June	30
March	31
May	31
November	30
October	32
September	60

From Table 1, we found the year 2001 to have 30 days with no records. 2011 had 120 missing records. Similarly, Table 2 shows missing days aggregated by months from 1991 to 2011. April had 30 days with no records. October has 32 missing records.

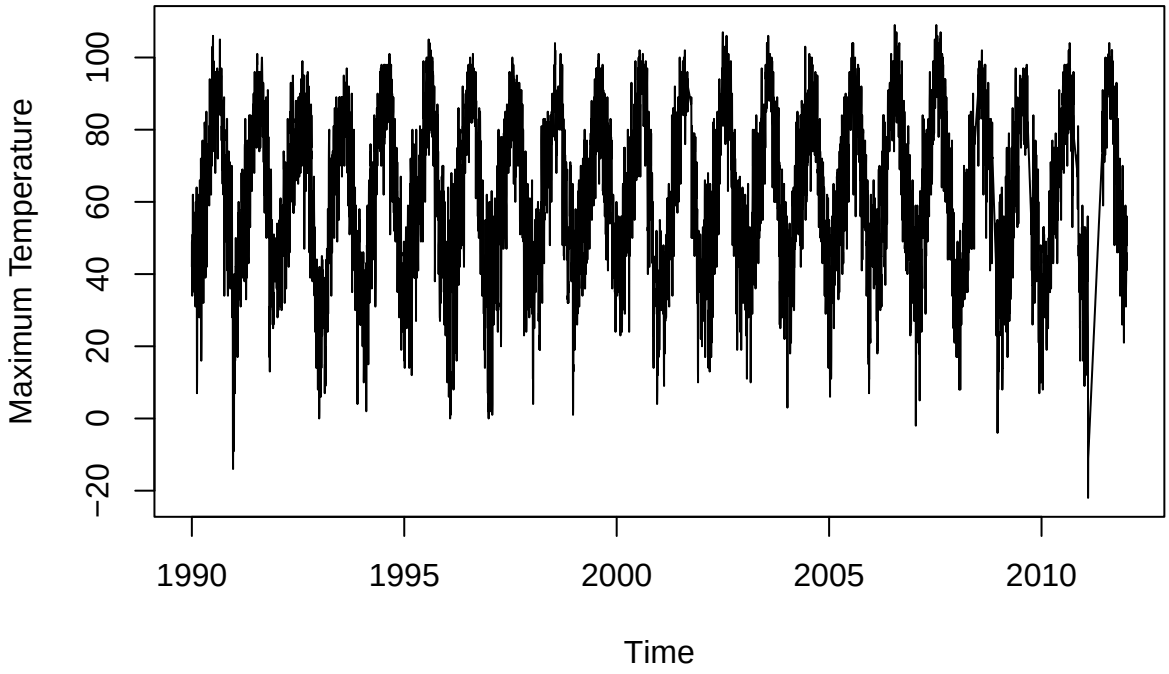

Figure 1: Time Series Plot of the Maximum Temperature

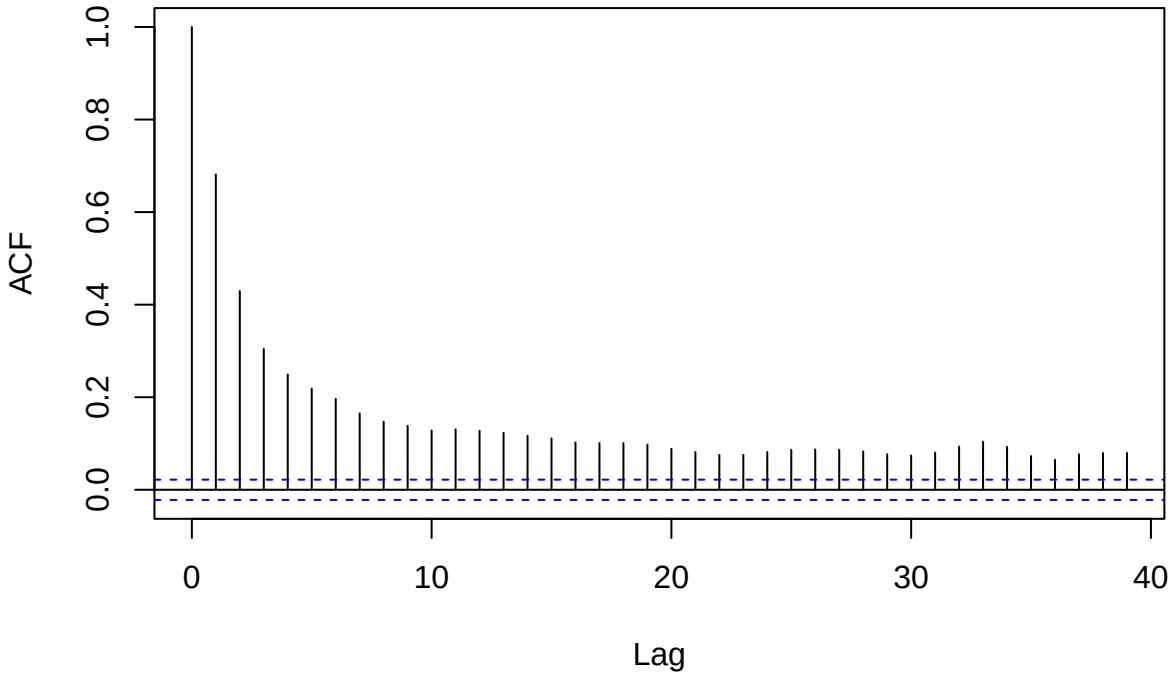

Figure 2: Correlogram Plot of the Residuals of the best-fitting harmonic model

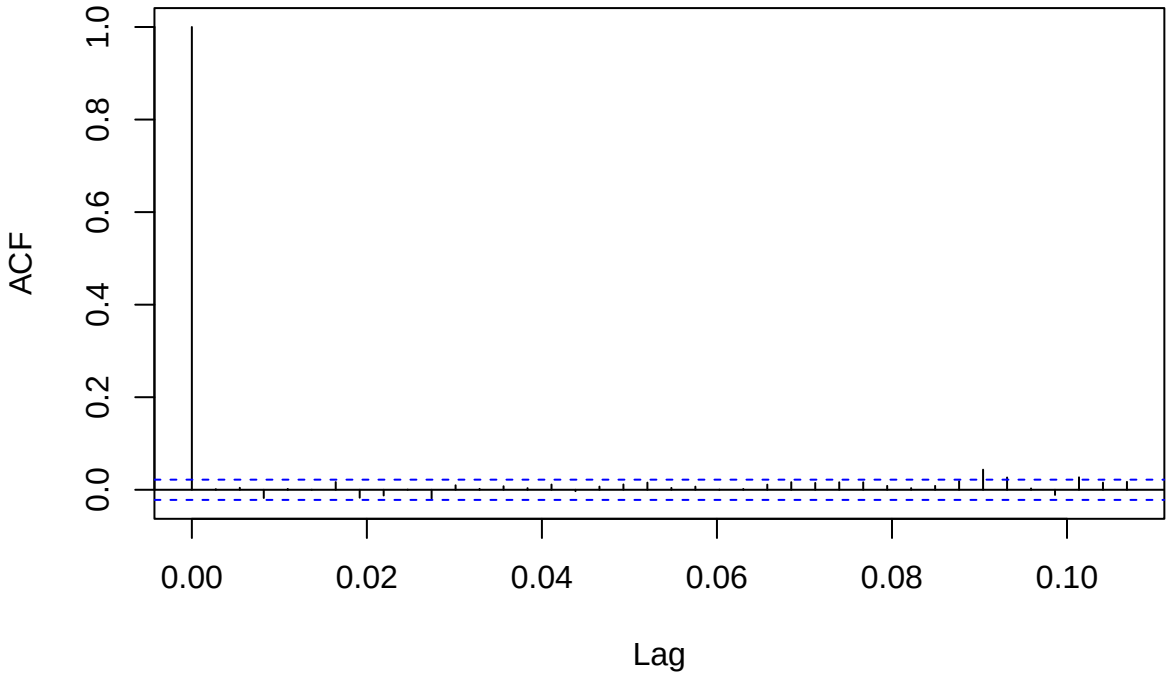

Figure 3: Correlogram of the Residuals of the best-fitting ARMA model

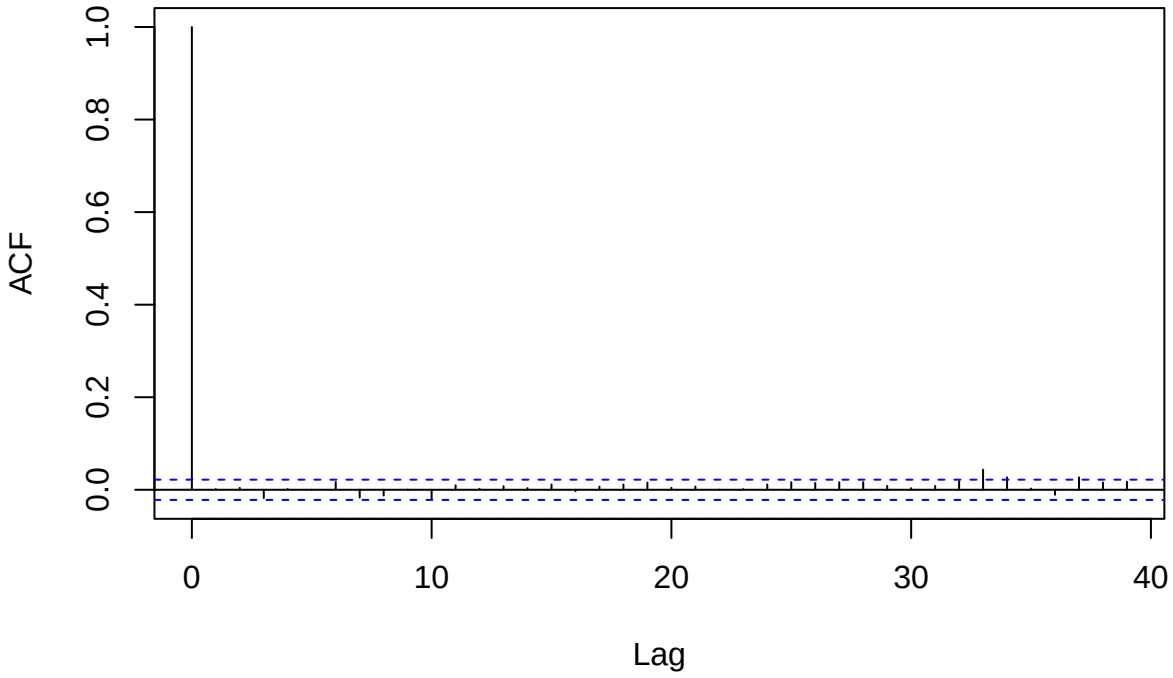

Figure 4: Correlogram of the residuals from the best-fitting SARIMA model

Table 3: AIC Comparison

Model	Harmonic	ARMA	ARIMA	SARIMA	GARCH
AIC	63450.92	58627.91	58748.77	58665.82	88536.23

Contributions

Theophilus Anim Bediako : missing value estimation, harmonic model and compilation of the paper

Michael Abalo : built arma model, arima model

Schuylar Harris Donovan : garch model, sarima model

Reference

1. Bishop, G., & Welch, G. (2001). An introduction to the kalman filter. Proc of SIGGRAPH, Course, 8(27599-23175), 41

2. Durbin,J., & Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. Biometrika, 89(3), 603-616

3. https://www.econ.cam.ac.uk/people-files/faculty/mw217/pdf/mispapnw.pdf

4. https://stats.stackexchange.com/questions/140990/using-kalman-filters-to-impute-missing-values-in-time-series

5. Metcalfe, A. V., & Cowpertwait, P. S. (2009). Introductory time series with R. Springer-Verlag New York.