STA 137 Final Project

Yesuunee Erdenebat ()

Yi Zhu ()

Introduction

In this project we worked on time series data table of each year's temperature anomaly to years from 1850-2019. Anomalies are critical in indicating the deviation from baseline temperature. The main objective of the project is to study the behavior of the given data table, forecast the trend and generate prediction for the future anomalies in the temperature. Furthermore, the main problem to be solved in the project is to find the best fit model to produce a quality forecasting for the *TempNH* data. The data is obtained from Climate Research Center, University of East Anglia, UK. By studying this time series data on temperature anomalies of the past years, we can predict the future anomaly of the temperature and make a informed decision based on such conclusion.

Materials and Methods

1. Data description

This data set contains annual temperature anomalies (1850-2019) for the northern hemisphere. There are two variables in this data set. Variable *year* refers to the period of time recorded from 1850 to 2019. Values of the other variable *Temp.NH* [1]are calculated based on the weighted average of all the non-missing, grid-box anomalies in the northern hemisphere, which describe the temperature differences of a corresponding year in comparison to the baseline. A positive anomaly indicates that the observed temperature is warmer whereas a negative anomaly indicates the observed temperature is cooler than the baseline.

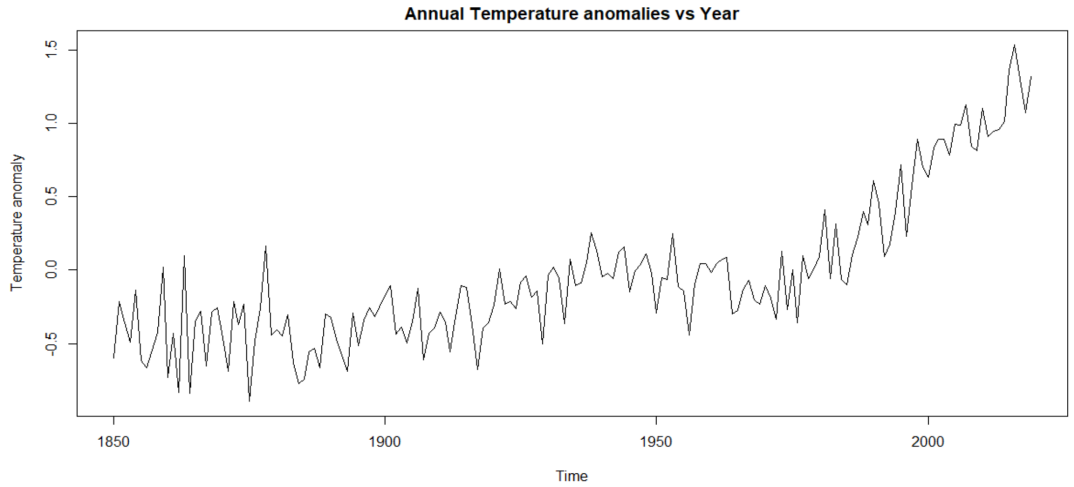

Fig.1 Annual Temperature anomalies vs Year

[1] Details can be found at https://crudata.uea.ac.uk/cru/data/temperature/#datdow

According to fig.1, there is a trend showing that as time progresses, the overall temperature anomalies increases. In addition, we can observe that there are fluctuations along the trend. Therefore, we conclude that this is a time series dataset with trend and we can implement time series analysis techniques.

2. Methods

2.1 Data transformation

We use box-cox data transformation to decrease the fluctuation of the trend over time of the *TempNH* data. (Unless necessary, we will not use the transformed data.)

2.2 Estimation of Trend

Loess is a point by point method of regressing linear regression model to each time point in a given interval to determine trend. In our case, the time interval is from 1850-2019. The temp anomaly is the linear regression model.

2.3 Estimation of Rough

Check for stationarity,

Checking for Stationarity of the Rough indicates that the time series properties will remain stable when constructing it. For our data *TempNH,* we would plot the data to test for stationarity. The assumptions for stationary rough series is:

1. The graph of the rough oscillates around same value Mu over time.
2. Variance is the same over time
3. Correlation over the interval of lag(j) do not change over time.

If the Rough part is not stationary, we cannot do further forecasting on the data, specifically ACF and PACF cannot be modeled if the sequence is not stationary.

2.4 ACF PACF

ACF plot (complete auto-corrolation function) shows the relationship between observation at current and previous time marks while considering trend and stationarity into account. ACF plot is constructed to determine if the data should be modeled by MA model. Only the lags with significant {et} will be modeled.

PACF plot (partial auto-corrolation function) reveals the relationship between two time spots without considering their relations to previous spots. PACF plot determines which lags of {et} are significant to build a AR model on.

Periodogram(and its smoothed version)

After building and fitting the model we use the spectral analysis to make a linear combination to find the disguised cyclicity. It could be hidden daily, monthly or yearly period. The way to build periodogram plot is:

1. Detrend the data to get the stationary model
2. Finding out the relationship between frequency and variability is the pinpoint of periodogram. Some frequencies have higher impact on variability. The function for the correlation between frequency and variability is called spectral density function (frequency of the sharpest peak).

Smoothing:

1. We filter the Periodogram we obtained in the first part to get rid of the white noise. By filtering we are obtaining a plot that is easy to read but also still has its major details
2. By using the special function we would connect the spectral density to out stationary model
3. From the equation we make the preliminary identification of the ARMA model.

2.5 Residuals

We fit the preliminary identification model to PACF and ACF and examine for significance.

2.6 Model Selection

AIC is a criterion for fitting a model and it chooses best fit model relative to the others it generates. Lowest AIC indicate the best model amongst the others. By using AIC criterion we find the best ARIMA model as our final model.

2.7 Spectral analysis

Plot Spectral Density of the final model of ARIMA

Spectral density is a function of the relationship between frequency and variability. Sharp peaks in the graphs are the spectral density.

Smoothed periodogram indicates the major frequency that is affecting the graph.

By using R we calculated the Spectral Density from Periodogram and

smoothed periodogram.

2.8 Model evaluation of the final model

We used ACF , PACF and smoothed periodogram of the final model to evaluate it.

2.9 Refit the final model

We used the forecasting function from the R function(forecast) to refit the final.

Results

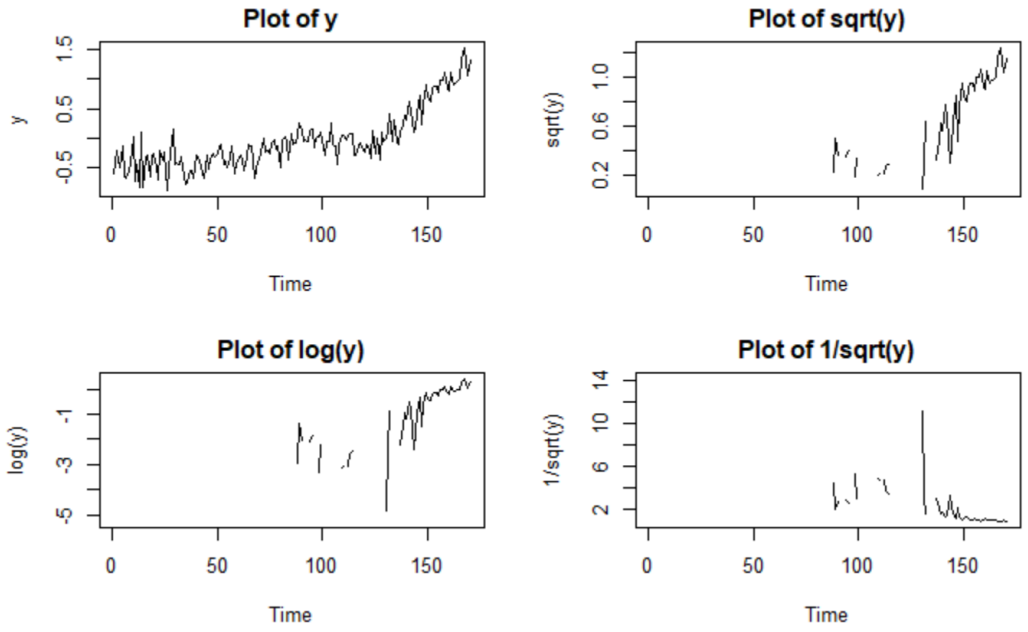

Fig.2 Box-cox transformation results

Since temperature difference ranges from -1 to 2, we did not obtain useful results from box-cox transformation on the original data. Additionally, R^2 for $\{Yt\}$ is 0.8699, indicating a reasonable fit. For the further analysis, we used the original dataset because data transformation did not help significantly.

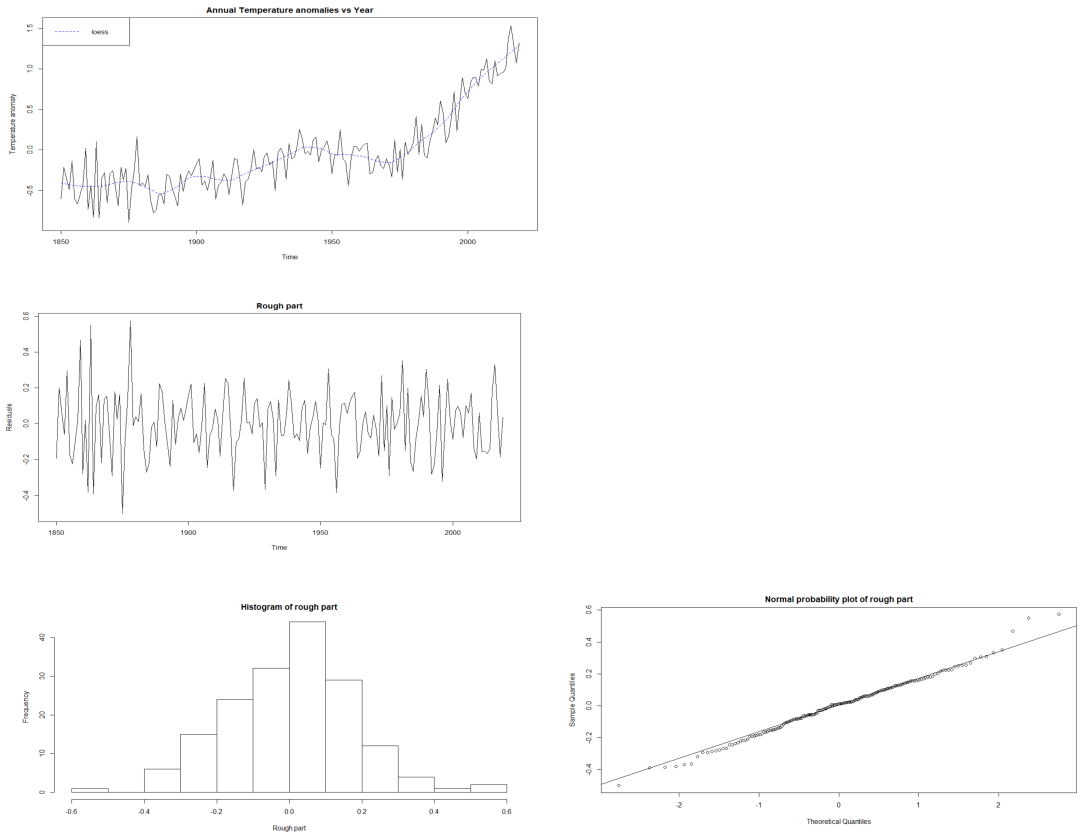

Fig3. Estimation of trend and rough part

As shown in fig.3, we fit a loess trend for this dataset and plot estimated Xt against time for loess model. Plot of rough part tells that the mean of Xt is approximately zero and variance of Xt (the size of fluctuations) does not change over time. Histogram and normal probability plot of rough part illustrate that residuals are normally distributed around 0. To sum up, graphical results suggest this is a (weak) stationary time series.

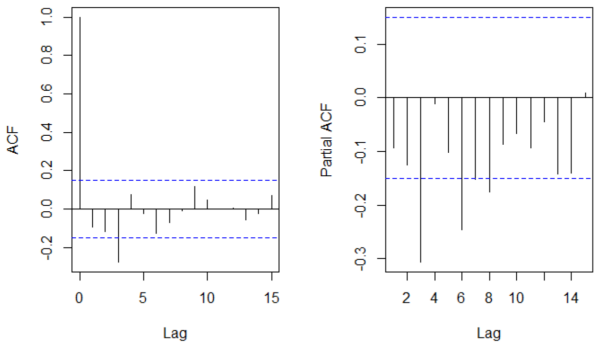

Fig4. ACF and PACF plot

For ACF plot we can see that there is a significant cut off at lag 3 after lag 1 while for PACF plot there are several significant values. Combined the output of ACF and PACF plots, we believe MA (3) or ARMA (0,0,3) might be a good model for the

rough part. (PACF interpretation)

Fit MA (3) model, residuals and properties

ARMA(p,0,q)	*q = 0*	*q = 1*	*q = 2*	*q = 3*
p = 0	-97.74038	-97.90062	-125.4216	-145.2461
p = 1	-97.22009	-134.3176	-135.4379	-144.5753
p = 2	-97.84938	-137.61182	-151.4208	-142.5875
p = 3	-112.43758	-139.99276	-140.1912	-140.5918

Fig5. AIC table for model selection

Here we computed AIC for ARMA(p,0,q), p = 1,2,3 and q = 1,2,3. From the AIC table, MA (3) has a low AIC value of -145.2461. However, ARMA (2,0,2) has the lowest AIC value in all the models.

Conclusion and discussion

To conclude, our group thinks that the model fit best to the data. From the analysis of the temperature vs time plot we found that the anomalies of the temp will have a trend of increasing. Periodograms indicate that, there is a certain frequency of periodicity in the detrended data. The graph shows that some frequency has higher influence on the variances. Lastly, we predicted the last six years temp of the anomalies. Fit was hard to identify but we think it is a reasonable fit for the model.

The ARMA mo