STA137 Final Project

Jasper Tsai: 916342467, Sirapat Watakajaturaphon: 920226951

2022-12-01

Contents

1 Introductions:

We are given a data set from the Climate Research Center, University of East Anglia, UK. The data contains annual temperature anomalies (a departure from a reference value or long-term average) from 1850 to 2021. Note that a positive anomaly means that the observed temperature was warmer than the reference value and negative anomaly means that the observed temperature was cooler than the reference value. We are interested in creating a model that is capable of forecasting future anomaly. It is important to see the trajectory of these future anomalies because anomalies describe how climate is changing over larger areas more clearly than temperatures themselves.

2 Materials and Methods:

The given data has annual temperature anomaly value from 1850 to 2021. This is a time series data because the data is recorded over consistent intervals of time. The time series plot shows that there is a trend that we can investigate to create a forecast and that the data is not normally distributed.

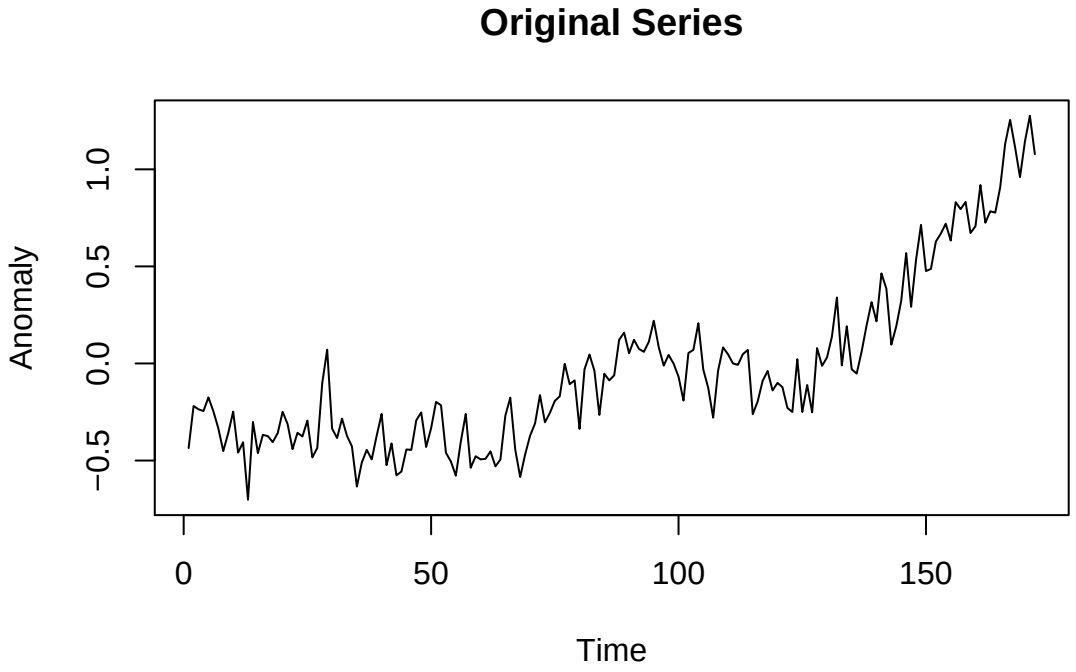
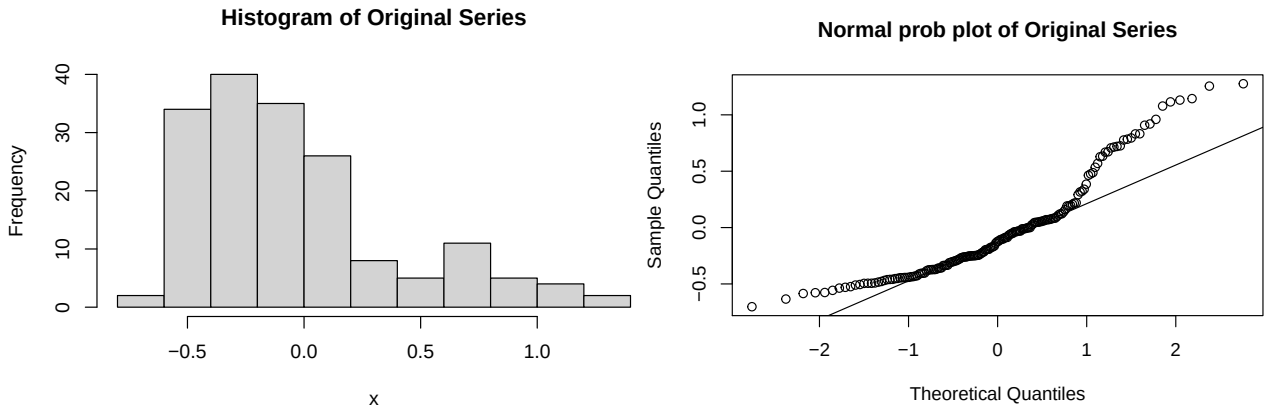


Figure 1: Plot of Original Series



(a) Histogram (b) NormalQQ

Figure 2: Diagnostics for Original Series

We will approach finding the appropriate model using *ARIMA Modeling* on the entire data.

2.1 Finding Model with entire data

In this section we use the method *ARIMA modeling* to approach our interest of creating a forecasting model for future anomaly . The first step is to look at the first difference using the following formulation:

$$X_t = \nabla Y_t = Y_t - Y_{t-1}$$

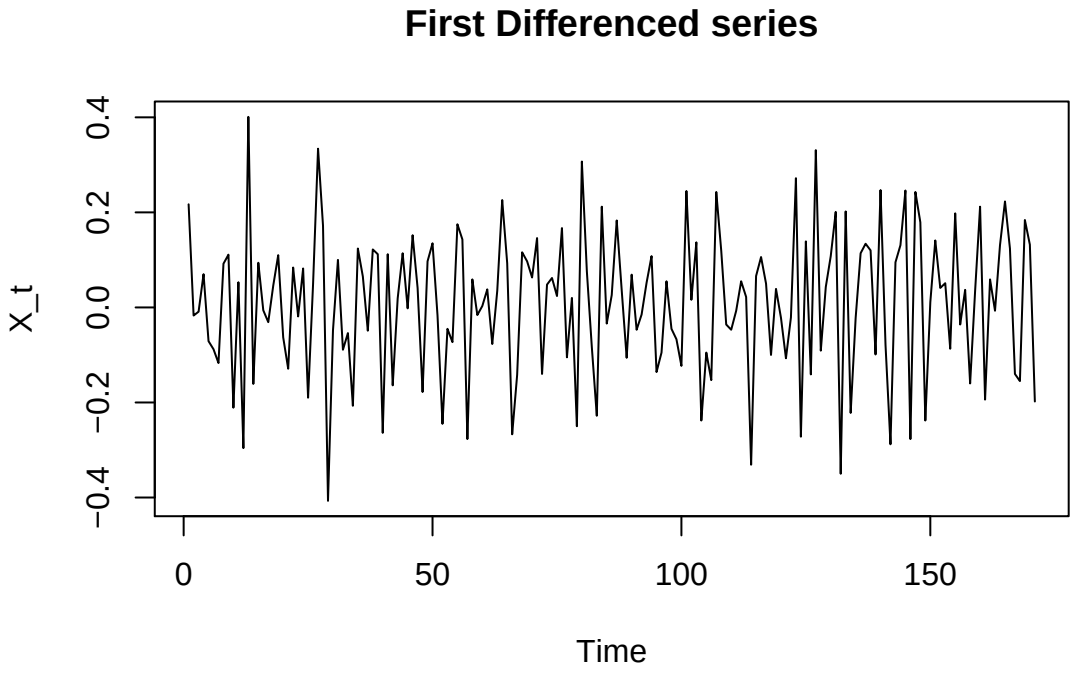
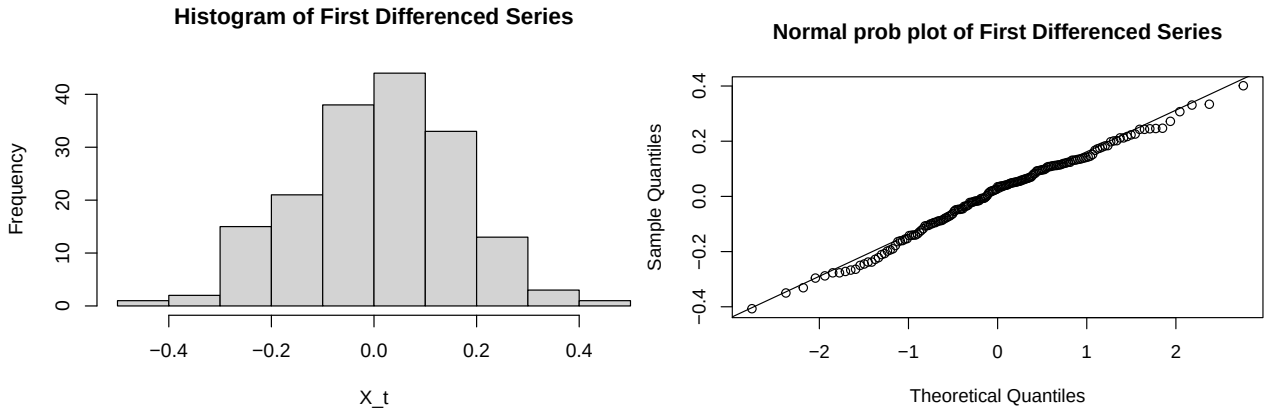


Figure 3: Plot of First Differenced Series

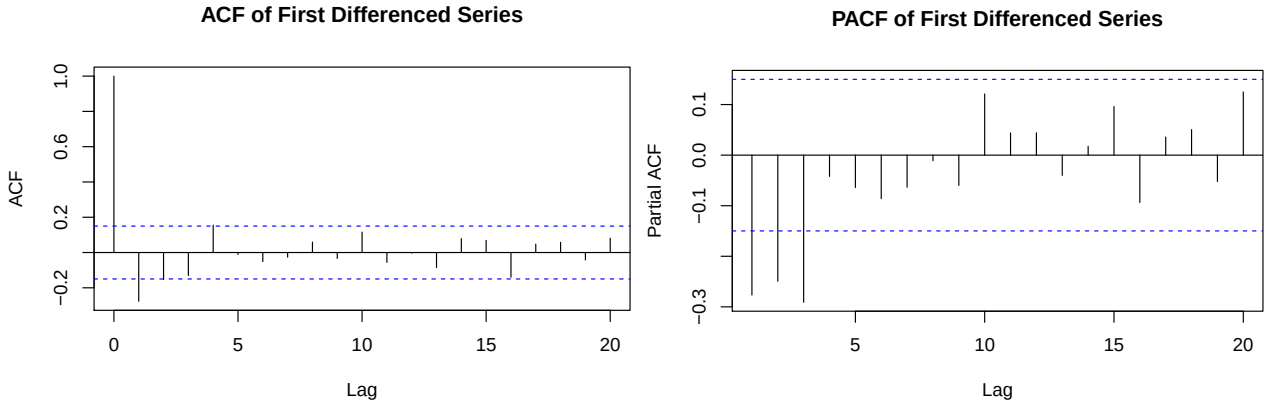
Now the trend seems to be gone and the first difference X_t looks stationary. The diagnostics now shows first difference series is normally distributed.



(a) Histogram (b) NormalQQ

Figure 4: Diagnostics for First Difference Series

We will proceed to examine the ACF and PACF plots for the differenced series:



(a) ACF (b) PACF

Figure 5: ACF and PACF Plots of First Differenced Series

From the ACF plot of the first difference, there is a cutoff after lag(1) and shows oscillating behavior. Our best guess for an MA model is MA(1).
We see that the PACF plot of first difference becomes 0 after lag(3) and shows minor oscillating behavior. Our best guess for an AR model is AR(3).
From initial inspection of the ACF and PACF plot we cannot rule out the possibility of an ARMA model. Our best guess for an ARMA model is ARMA(3,1).

We will fit ARMA(3,1) on the first difference series which is equivalent to ARIMA(3,1,1) if using the original anomaly values.
We will call this **model1** and examine the residuals and their properties to make sure our preliminary model fit is appropriate.

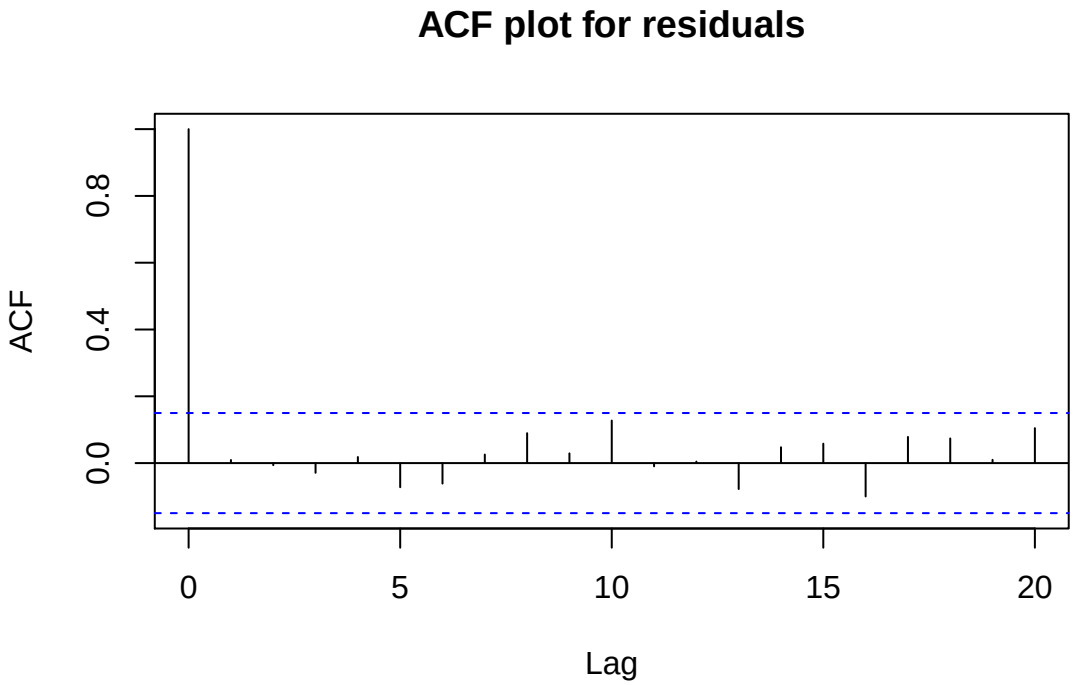


Figure 6: ACF Plot for ARIMA(3,1,1) Residuals

We see that residuals for ARMA(3,1) of first difference is a reasonable model for the data because ACF plot indicate that the residuals are iid and we see that after lag 0, the rest are inside the the bands. Residuals approximately behave like a white noise.

We will now use the AIC model selection criterion to check whether or not our preliminary fit of ARMA(3,1) for the first difference model is indeed the best model or if there is a better model based on the AIC criterion.

Our ARMA(p,q) for first difference model is the same as ARIMA(p,1,q) for original anomaly values.
We will consider the following 16 models $0 \leq p \leq 3, 0 \leq q \leq 3$, where p is the AR order and q is the MA order.

This is the AIC criterion table. We will try to select the most suitable model base on the lowest AIC value from the table.

Table 1: AIC for first difference series

	q=0	q=1	q=2	q=3
p=0	-0.9229414	-1.094005	-1.119878	-1.108348
p=1	-0.9922384	-1.114102	-1.108237	-1.109690
p=2	-1.0455561	-1.114930	-1.112556	-1.114580
p=3	-1.1233958	-1.116518	-1.110751	-1.103335

From the AIC criterion, it is suggested that ARIMA(3,1,0) or [ARMA(3,0) or AR(3) for the first difference series] is the best model. So we will fit this model and name it **model2**.

Again we will check ACF plot of residuals and we see that that they are iid. Residuals approximately behave like a white noise.

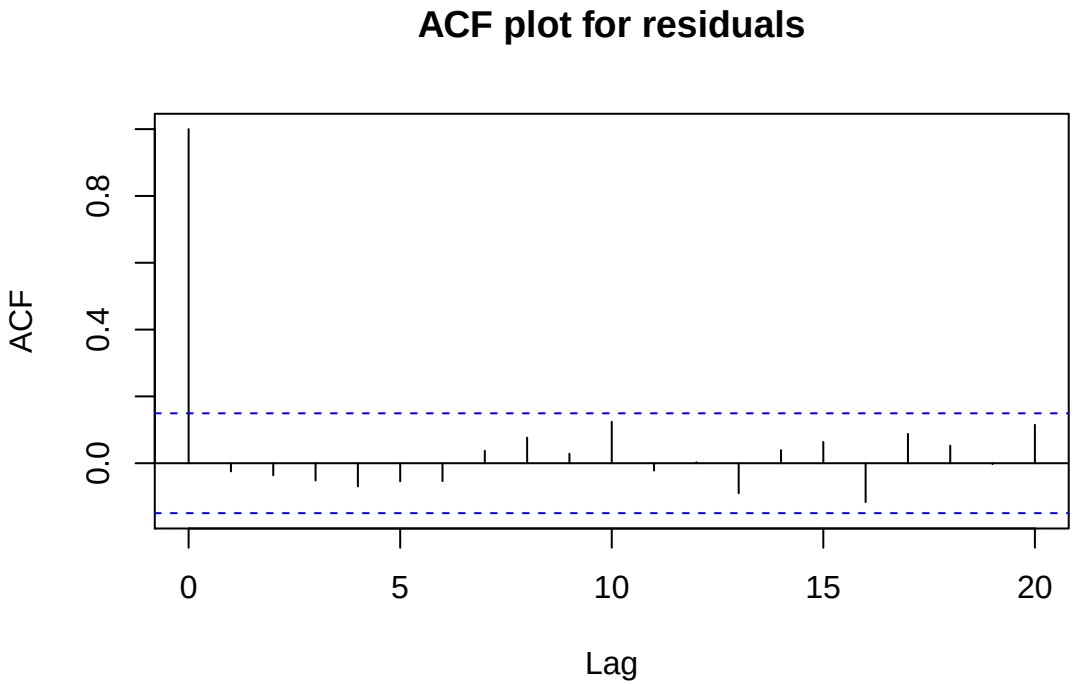


Figure 7: ACF Plot for ARIMA(3,1,0) Residuals

We will do a formal test for checking if the sequence is i.i.d.
$H_o : \rho(1) = ... = \rho(h) = 0$
H_a: at least one of $\rho(1), ..., \rho(h)$ is nonzero

```
##
##  Box-Ljung test
##
## data:  model2$residuals
## X-squared = 7.032, df = 10, p-value = 0.7224
```

Test statistic: 7.0319746 and the p-value: 0.722423

Given that the p-value is large, we fail to reject H_o and conclude that the sequence is i.i.d.

This suggest that our model of ARIMA(3,1,0) from AIC criterion is also valid.
Proceeding with **model2** [ARIMA(3,1,0], we have the following estimated parameters and the standard errors:

Table 2: Estimated Parameters for ARIMA(3,1,0)

	Parameter Estimate
ar1	-0.4113455
ar2	-0.3429713
ar3	-0.2837367

Table 3: Standard Errors for ARIMA(3,1,0)

	Standard Error
ar1	0.0738018
ar2	0.0758651
ar3	0.0738616

We will check the fit of **model2**[ARIMA(3,1,0)] by plotting the fitted side by side with the observed data

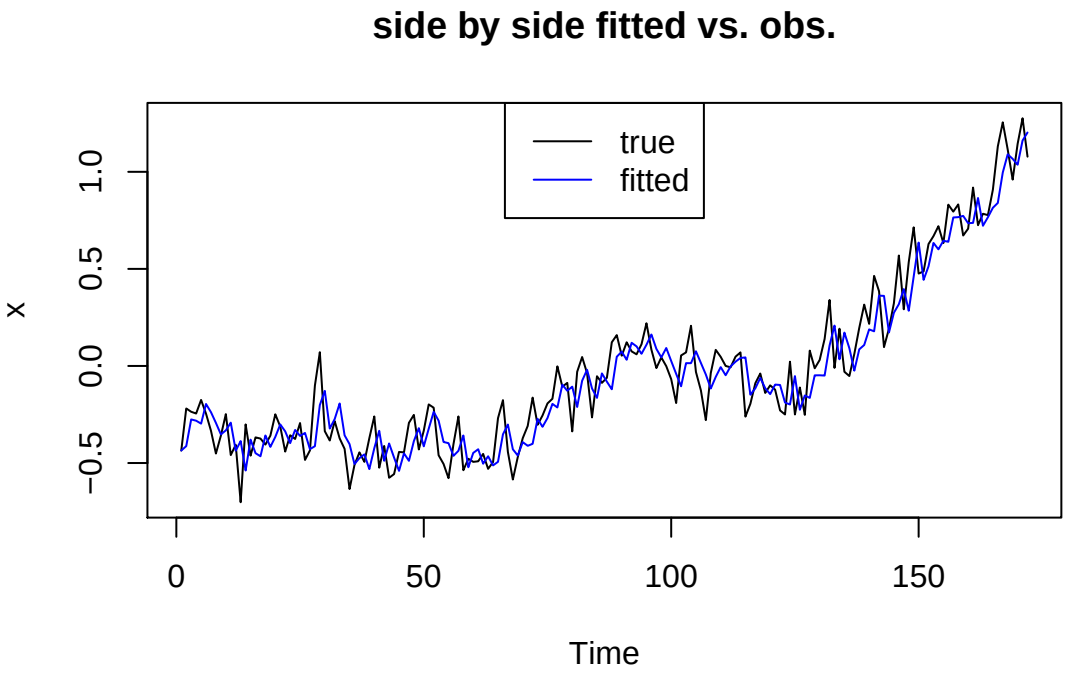


Figure 8: ARIMA(3,1,0) fitted vs. observed

2.2 Forecasting temperature anomalies for the last 6 years.

Assume that now we are in 2015 and we would like to forecast the temperature anomalies for 2016~2021. We will refit our final model from the first differenced series collected from the previous section [AIC criterion] and refit it using all data except for the last 6 years. We will then compare it to the other the method of *Modeling Trend by estimating the rough* (using all data except for last 6 years) and observe which method is more ideal.

2.2.1 ARIMA Modeling (forecast 2016~2021)

We will refit the previously selected ARIMA(3,1,0) onto the data with last 6 years removed, call it **model3** to do forecasting on 2016~2021.

Table 4: Forecast using First Difference method (2016~2021)

	forecast value
2016	0.9931122
2017	0.9391939
2018	0.9459326
2019	1.0004318
2020	0.9901110
2021	0.9740973

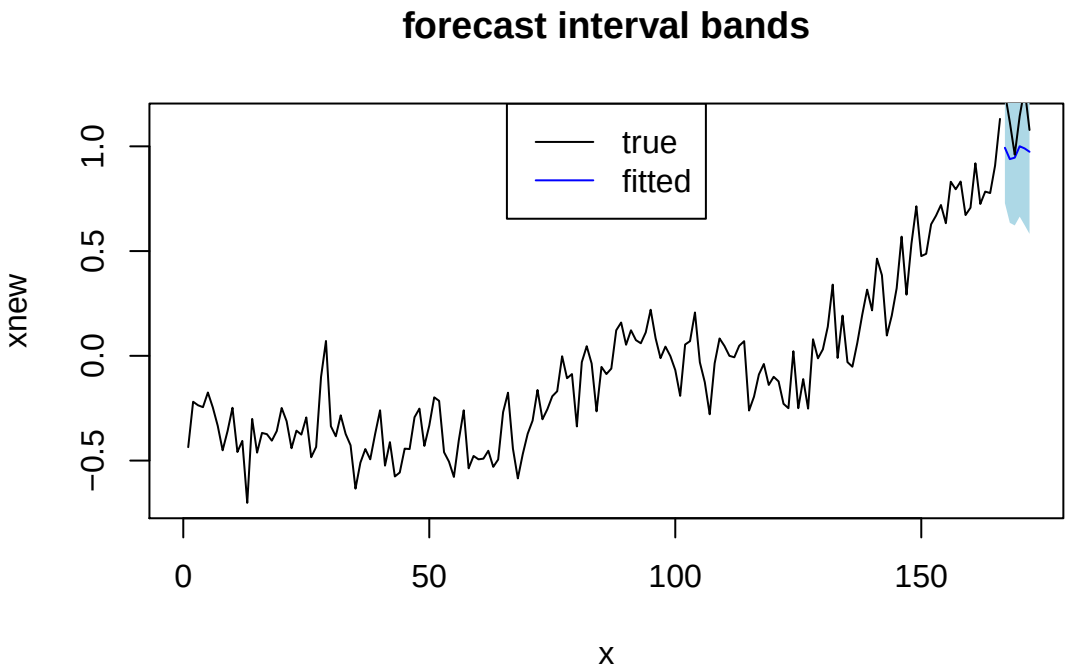


Figure 9: 2016 to 2021 fitted vs. observed with forecast interval bands

It is desirable for us to compare forecasting by different methods, now we will attempt *Modeling Trend by estimating the rough*

2.2.2 Modeling Trend by estimating the rough (forecast 2016~2021)

Using $Y_t = m_t + X_t$ where m_t is smooth/trend, X_t is the rough, and Y_t is observed value from the data.

After using the Cubic Spline method to estimate m_t , we will plot the estimated trend along with the time series in the same graph.

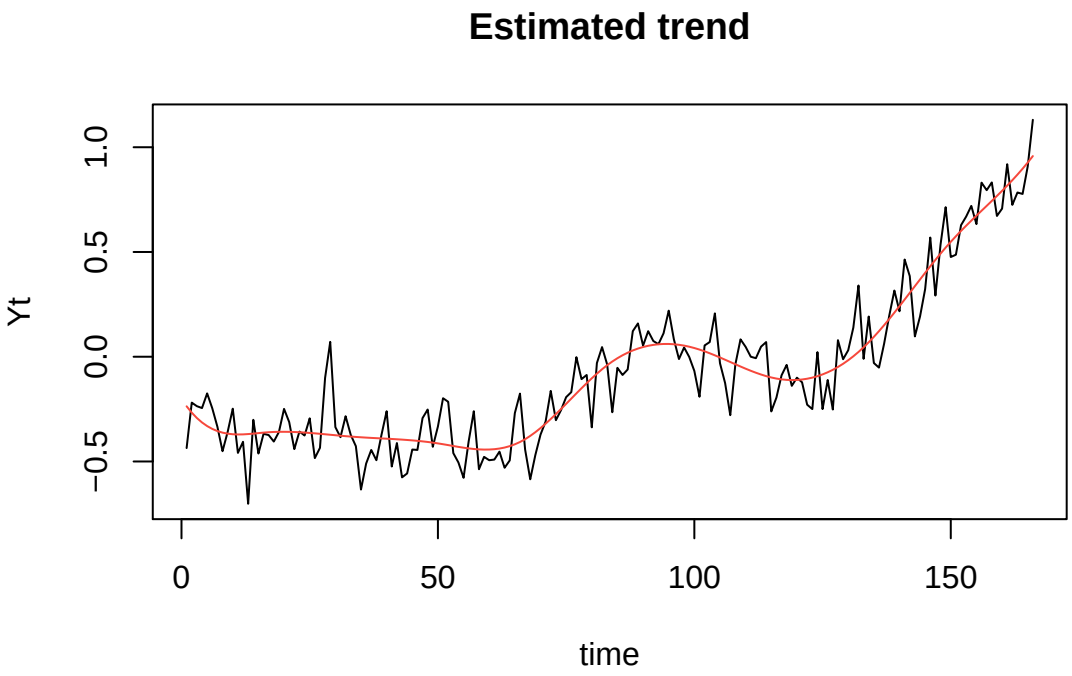


Figure 10: estimated trend vs. observed

The fit looks appropriate and we will proceed to estimate the rough part using the following: $\hat{X}_t = Y_t - \hat{m}_t$

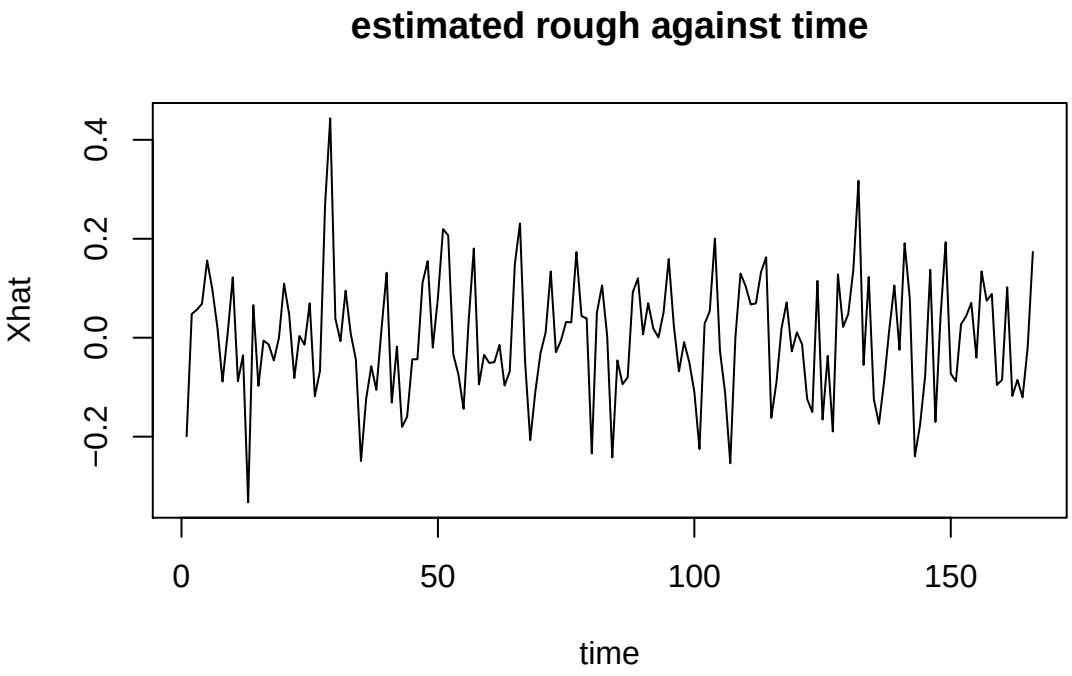
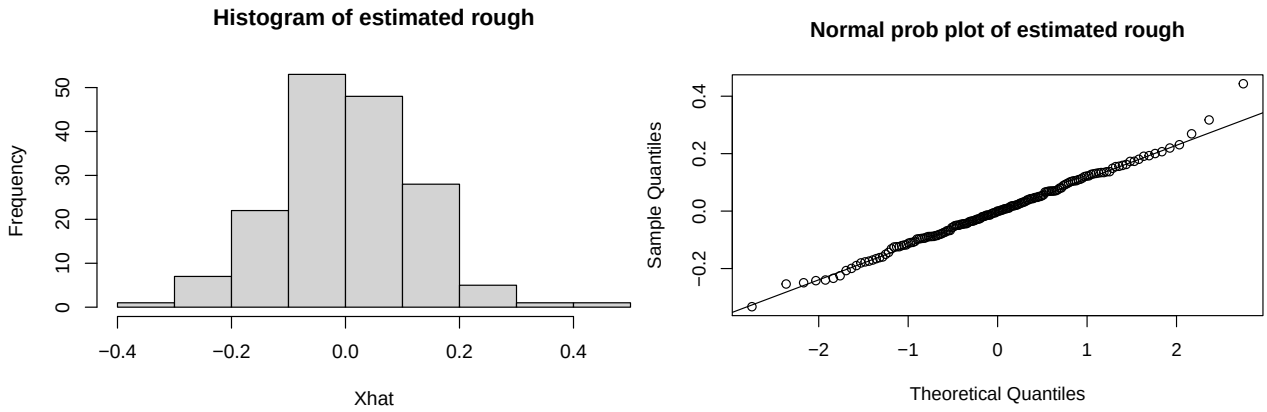


Figure 11: estimated rough vs. time

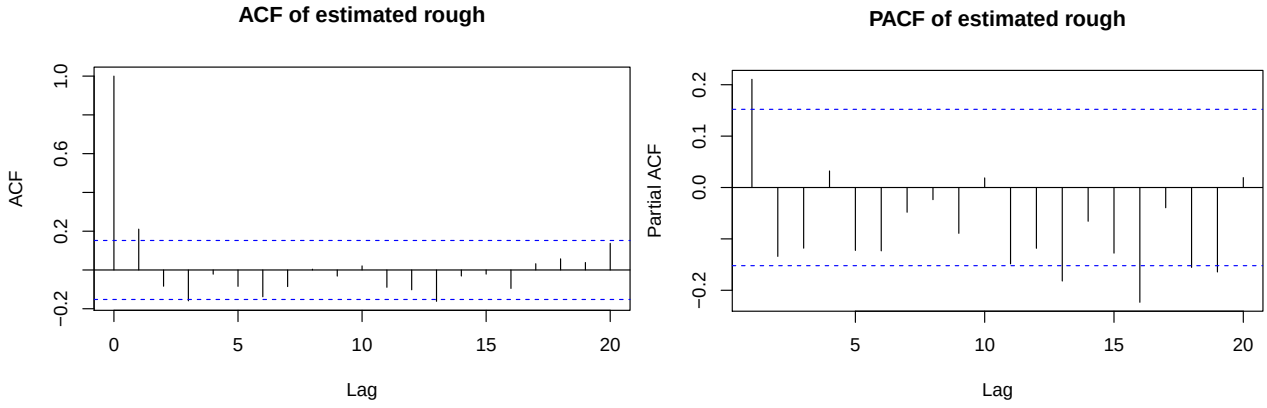
We see that the estimated rough now appears stationary. The diagnostics now shows estimated rough is normally distributed.



(a) Histogram (b) NormalQQ

Figure 12: Diagnostics for estimated rough

We will now obtain the ACF and PACF plot for the estimated $\hat{X}_t$



(a) ACF (b) PACF

Figure 13: ACF and PACF Plots of estimated rough

From the ACF plot of the estimated rough, there is a cutoff after lag(1) and shows oscillating behavior. Our best guess for an MA model is MA(1).
We see that the PACF plot of the estimated rough becomes 0 after lag(1) and shows minor oscillating behavior. We also notice that at lag(16) it passes the bands. Our best guess for an AR model is AR(1).
From initial inspection of the ACF and PACF plot we cannot rule out the possibility of an ARMA model. Our best guess for an ARMA model is ARMA(1,1).

We will consider the following 16 models $0 \leq p \leq 3, 0 \leq q \leq 3$, where p is the AR order and q is the MA order.

This is the AIC criterion table. We will try to select the most suitable model base on the lowest AIC value from the table.

Table 5: AIC for estimated rough

	q=0	q=1	q=2	q=3
p=0	-1.376682	-1.419367	-1.407399	-1.434691

	q=0	q=1	q=2	q=3
p=1	-1.411067	-1.407348	-1.408202	-1.535528
p=2	-1.418358	-1.543668	-1.415364	-1.525949
p=3	-1.420840	-1.413255	-1.530325	-1.513927

From the AIC criterion, it is suggested that ARMA(2,1) is the best model. So we will fit this model and name it **model4**.

Again we will check ACF plot of residuals and we see that that they are iid. Residuals approximately behave like a white noise.

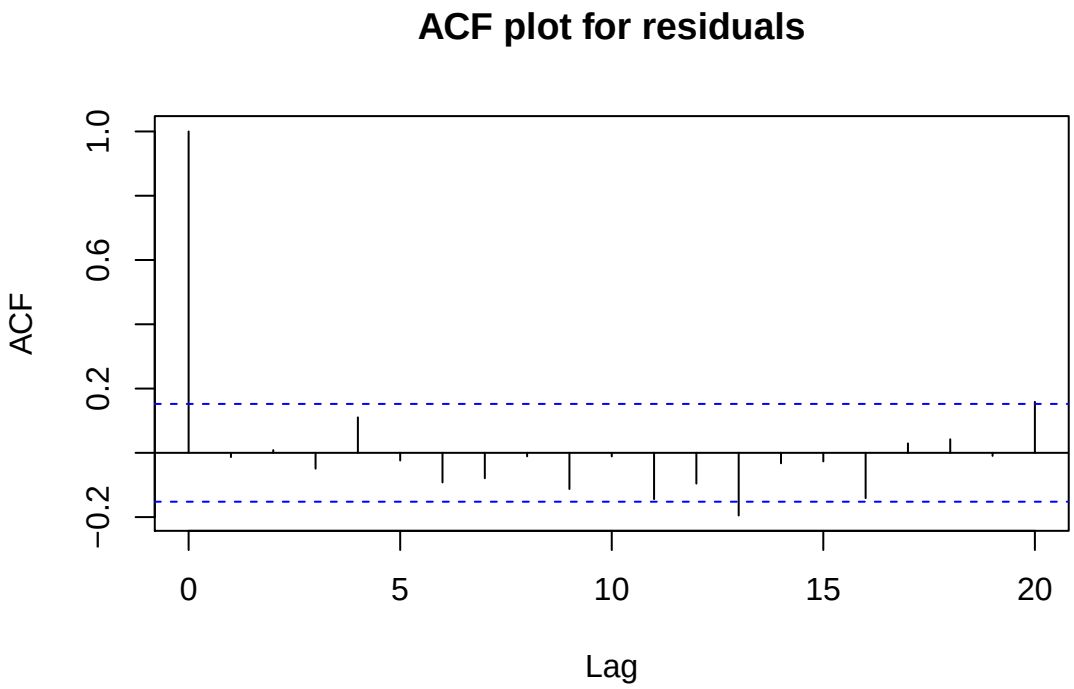


Figure 14: ACF Plot for ARMA(2,1) Residuals

Table 6: Forecast using Modeling Trend by estimating the rough (2016~2021)

	forecast value
2016	1.048311
2017	1.025480
2018	1.036576
2019	1.064226
2020	1.098345
2021	1.133867

3 Results:

Now we will check the forecasts with the two different methods and compare them with the actual observed values for 2016 to 2021.

Looking at *Table 4* where we used ARIMA modeling First difference method to forecast and *Table 6* where we forecast using Modeling Trend by estimating the rough, we see that the values in *Table 6* are slightly greater. The below Forecast results for 2016 to 2021 shows this. Note: In the graph, *Method 1* is First Difference ARIMA modeling forecast and *Method 2* is forecast using Modeling Trend by estimating the rough.

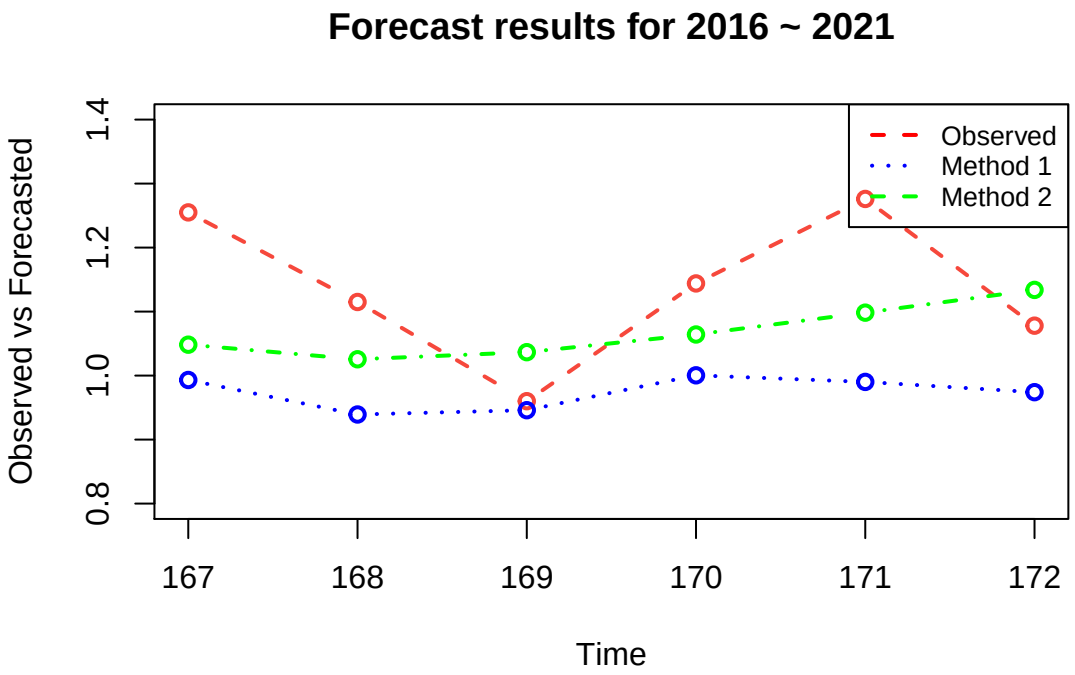


Figure 15: Forecasting results vs. observed for 2016 to 2021

From *Figure 15* we see that the values of *Method 2* is more similar to the Observed values for 2016 to 2021 than *Method 1*. Which suggest that *Method 2* is better at forecasting future annual temperature anomalies.

4 Conclusion and Discussion:

From our analysis, both methods does a good job forecasting. We did conclude that *Method 2* of forecast using Modeling Trend via cubic spline and estimating the rough with ARMA(2,1) is slightly better than *Method 1*, forecasting using First difference under ARIMA(3,1,0) because the values for the *Method 2* aligns closer to the values of the observed when forecasting 2016 to 2021 with data that goes from 1850 to 2015. We can possibly improve our forecast by exploring more ways to estimate trends to model the rough for *Method 2* or refit a new model with data from 1850 to 2015 when trying to forecast 2016 to 2021 (although this might not make a difference in the selected model).

5 Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
require(readxl)
require(forecast)
require(astsa)
require(Hmisc)
data = read_excel("TempNH_1850_2021.xlsx")
data
#Grab anamoly values
x <- data[,2]
plot.ts(x, main = "Original Series")
x<-unlist(x)
hist(x, main = "Histogram of Original Series")
qqnorm(x, main = "Normal prob plot of Original Series")
qqline(x)
X_t<-diff(x,1)
plot.ts(X_t, main = "First Differenced series")
hist(X_t, main = "Histogram of First Differenced Series")
qqnorm(X_t, main = "Normal prob plot of First Differenced Series")
qqline(X_t)
acf(X_t, lag.max = 20, main ="ACF of First Differenced Series")
pacf(X_t,lag.max = 20, main = "PACF of First Differenced Series")
model1 = arima(X_t, order = c(3,0,1))
acf(model1$residuals, lag.max = 20, main="ACF plot for residuals")

AIC<-matrix(0,4,4)
for (i in 1:4){
  for (j in 1:4){
    AIC[i,j]<-sarima(x,p=i-1,d=1,q=j-1,details=FALSE)$AIC
  }
}
#AIC
#Get smallest value
#which(AIC == min(AIC), arr.ind = TRUE)
rownames(AIC) <- c("p=0", "p=1", "p=2", "p=3")
colnames(AIC) <- c("q=0", "q=1", "q=2", "q=3")
knitr::kable(AIC,caption = "AIC for first difference series", align= "c")
model2 = arima(x,order=c(3,1,0))
acf(model2$residuals, lag.max = 20, main="ACF plot for residuals")
Boxtest = Box.test(model2$residuals, lag=10, type='Ljung-Box')
Boxtest
#model2$coef

knitr::kable(model2$coef,caption = "Estimated Parameters for ARIMA(3,1,0)", align= "c",col.names = "Para
# model2$var.coef##variance covariance matrix, so select diagonals to get variance
#sqrt(diag(model2$var.coef))##std.error = sqrt(variance)

knitr::kable(sqrt(diag(model2$var.coef)),caption = "Standard Errors for ARIMA(3,1,0)", align= "c",col.na
n <- length(x)
plot.ts(x, main="side by side fitted vs. obs.")

lines(1:n, fitted(model2), col = "blue")
```

```
legend("top", legend = c("true","fitted"), lty=c(1, 1), col = c("black","blue"))
#split data
xnew <- x[1:(n-6)]
xlast <- x[(n-5):n]
#fit
model3 <- arima(xnew,order = c(3,1,0)) #prediction
h <- 6
m <- n - h
fcast1 <- predict(model3, n.ahead=h)
upper1 <- fcast1$pred+1.96*fcast1$se
lower1 <- fcast1$pred-1.96*fcast1$se
fcast_Y <-as.data.frame(fcast1$pred)
rownames(fcast_Y) = c("2016","2017","2018","2019","2020","2021")

knitr::kable(fcast_Y,caption = "Forecast using First Difference method (2016~2021)", col.names = "foreca

#plot
plot.ts(xnew, xlim = c(0,n), xlab = "x", main = "forecast interval bands")
polygon(x=c(m+1:h,m+h:1), y=c(upper1,rev(lower1)), col='lightblue', border=NA)
lines(x=m+(1:h), y=fcast1$pred,col='blue')
lines(x=m+(1:h), y=xlast,col='black')

legend("top", legend = c("true","fitted"), lty=c(1, 1), col = c("black","blue"))
# fit the (cubic) spline trend
trend_spline = function(y, lam){ # Prof. Burman's spline trend function
    n   = length(y)
    p   = length(lam)
    rsq = rep(0, p)
    y   = sapply(y,as.numeric)
    tm  = seq(1/n, 1, by=1/n)
    xx  = cbind(tm, tm^2, tm^3)
    knot= seq(.1, .9, by=.1)
    m   = length(knot)
    for (j in 1:m) {
      u = pmax(tm-knot[j], 0); u=u^3
      xx= cbind(xx,u)
    }
    for (i in 1:p) {
      if (lam[i]==0) {
        ytran = log(y)
      } else {
        ytran = (y^lam[i]-1)/lam[i]
      }
      ft  = lm(ytran~xx)
      res = ft$resid ; sse = sum(res^2)
      ssto= (n-1)*var(ytran)
      rsq[i] = 1-sse/ssto
    }
    ii=which.max(rsq) ; lamopt=lam[ii]
    if (lamopt==0) {
      ytran = log(y)
    } else {
      ytran = y^lamopt
```

```
    }
    ft = lm(ytran~xx);
    best_ft = step(ft, trace=0)
    fit = best_ft$fitted ; res = best_ft$resid
    result = list(ytrans=ytran, fitted=fit, residual=res, rsq=rsq, lamopt=lamopt)
    return(result)
}
#estimating the trend of all data except last 6 points
Yt <- xnew
Yt<-unlist(Yt)
time = 1:(n-6)
mt = trend_spline(Yt,1) # lambde=1 means no transformation
## plot
plot(time,Yt,type = 'l', main='Estimated trend')
lines(time,mt$fitted,col=2)
Xhat =Yt-mt$fitted
plot(time,Xhat,main="estimated rough against time", type='l')
hist(Xhat, main = "Histogram of estimated rough")
qqnorm(Xhat, main = "Normal prob plot of estimated rough")
qqline(Xhat)
acf(Xhat, lag.max = 20, main ="ACF of estimated rough")
pacf(Xhat,lag.max = 20, main = "PACF of estimated rough")

AIC2<-matrix(0,4,4)
for (i in 1:4){
  for (j in 1:4){
    AIC2[i,j]<-sarima(Xhat,p=i-1,d=0,q=j-1,details=FALSE)$AIC
  }
}
#AIC2
#Get smallest value
#which(AIC2 == min(AIC2), arr.ind = TRUE)
rownames(AIC2) <- c("p=0", "p=1", "p=2", "p=3")
colnames(AIC2) <- c("q=0", "q=1", "q=2", "q=3")
knitr::kable(AIC2,caption = "AIC for estimated rough", align= "c")
model4 = arima(Xhat,order=c(2,0,1))
acf(model4$residuals, lag.max = 20, main="ACF plot for residuals")
#estimated trend for the last 6 entries
mthat = approxExtrap(x=1:(n-6), y=mt$fitted, xout = (n-5):n)
mthat = mthat$y
#estimated rough for the last 6 entries
xthat <- predict(model4, n.ahead=h)
xthat <- xthat$pred
#fitted values for the last 6 entries
Ythat <-mthat+xthat
Ythat <-as.data.frame(Ythat)
rownames(Ythat) = c("2016","2017","2018","2019","2020","2021")
knitr::kable(Ythat,caption = "Forecast using Modeling Trend by estimating the rough (2016~2021)", col.na
obs = xlast
plot((n-5):n, obs, type='b', col=2, lty=2, lwd=2, xlab='Time',
     ylab='Observed vs Forecasted', ylim = c(0.8,1.4), main = "Forecast results for 2016 ~ 2021")

lines((n-5):n, unlist(fcast_Y), type='b', col='blue', lty=3, lwd=2)
```

```
lines((n-5):n, unlist(Ythat), type='b', col='green', lty=4, lwd=2)


legend('topright', legend=c('Observed', 'Method 1',"Method 2"), col=c('red','blue','green'),
       lty=c(2,3), lwd=c(2,2), cex=0.8)
```