Lab 1: ARIMA models

Team member names: Nick Chambers (SAFS), Liz Elmstrom (SAFS), Maria Kuruvilla (QERM)

Data

We will use Bristol Bay Data. We want to use sockeye salmon returns for each river for each year.

Question your team will address

Our team decided to compare the accuracy of forecasts using best fit ARIMA models for Sockeye salmon in different rivers in Bristol Bay data. Our question is whether forecast accuracy is different for different rivers.

Method you will use

- fit ARIMA models (note you'll want to log the abundance data). You can fit other models in addition ot ARIMA if you want.
- do diagnostics for ARIMA models
- make forecasts
- test how good your forecasts or compare forecasts (many options here)

Initial plan

Describe what you plan to do to address your question. Note this example is with Ruggerone & Irvine data but your team will use the Bristol Bay data.

Example, "We will fit ARIMA models to 1960-1980 data on pink salmon in 2 regions in Alaska and 2 regions in E Asia. We will make 1981-1985 forecasts (5 year) and compare the accuracy of the forecasts to the actual values. We will measure accuracy with mean squared error. We will use the forecast package to fit the ARIMA models."

What you actually did

Example, "We were able to do our plan fairly quickly, so after writing the basic code, we divided up all 12 regions in Ruggerone & Irvine and each team member did 4 regions. We compared the accuracy of forecasts for different time periods using 20-years for training and 5-year forecasts each time. We compared the RMSE, MAE, and MAPE accuracy measures."

Diagnostics and preliminary exploration

Read the data

```
bb_data <- readRDS(here::here("Lab-1", "Data_Images",
                              "bristol_bay_data_plus_covariates.rds"))
head(bb_data)
```

```
##   brood_yr ret_yr  system fw_age o_age age_group       ret forecast.adfw
## 1     1957   1963 Igushik      2     3       2.3  9.213011            NA
## 2     1958   1963 Igushik      1     3       1.3 37.356730            NA
## 3     1958   1963 Igushik      2     2       2.2 14.002169            NA
## 4     1958   1964 Igushik      2     3       2.3 13.488280            NA
## 5     1959   1963 Igushik      1     2       1.2 75.778787            NA
## 6     1959   1964 Igushik      1     3       1.3 90.589305            NA
##   forecast.fri env_pdo  env_sst  env_slp env_upstr
## 1           NA -0.1125 6.079020       NA        NA
## 2           NA -0.1125 6.079020       NA        NA
## 3           NA  0.2250 6.674975 1012.555    6.2501
## 4           NA  0.2250 6.674975 1012.555    6.2501
## 5           NA  0.2250 6.674975 1012.555    6.2501
## 6           NA  0.2250 6.674975 1012.555    6.2501
```

Checking out the unique data groups

```
## colnames:  brood_yr ret_yr system fw_age o_age age_group ret forecast.adfw forecast.fri env_pdo env_s
```

```
## system (river):  Igushik Wood Nushagak Kvichak Naknek Egegik Ugashik
```

```
## age groups:  2.3 1.3 2.2 1.2
```

Retrieving a subset of the data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
subdata <- bb_data %>%
  group_by(system, ret_yr) %>%
  summarize(lntotal = log(sum(ret, na.rm=TRUE)))%>%
  filter(system == "Ugashik" || system == "Wood")
```

```
## `summarise()` has grouped output by 'system'. You can override using the
## `.groups` argument.
```

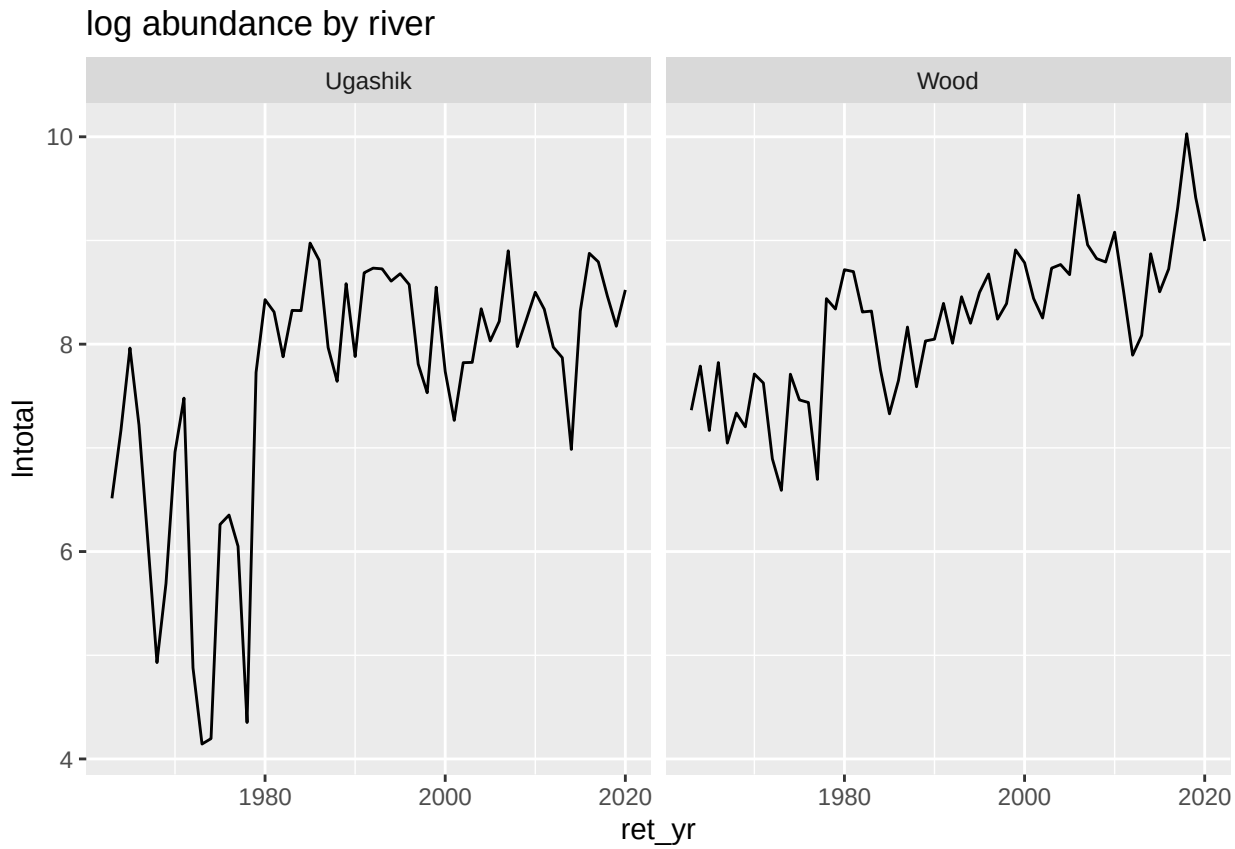
```
head(subdata)
```

```
## # A tibble: 6 x 3
## # Groups:   system [1]
##   system  ret_yr lntotal
##   <chr>    <dbl>   <dbl>
## 1 Ugashik   1963    6.51
## 2 Ugashik   1964    7.17
## 3 Ugashik   1965    7.96
## 4 Ugashik   1966    7.22
## 5 Ugashik   1967    6.08
## 6 Ugashik   1968    4.93
```

Plot the data

Plot the data and discuss any obvious problems with stationarity from your visual test.

```
plot1 <- ggplot(data = subdata, aes(x=ret_yr, y=lntotal)) +
  geom_line() +
  ggtitle("log abundance by river") +
  facet_wrap(~system)
plot1
```

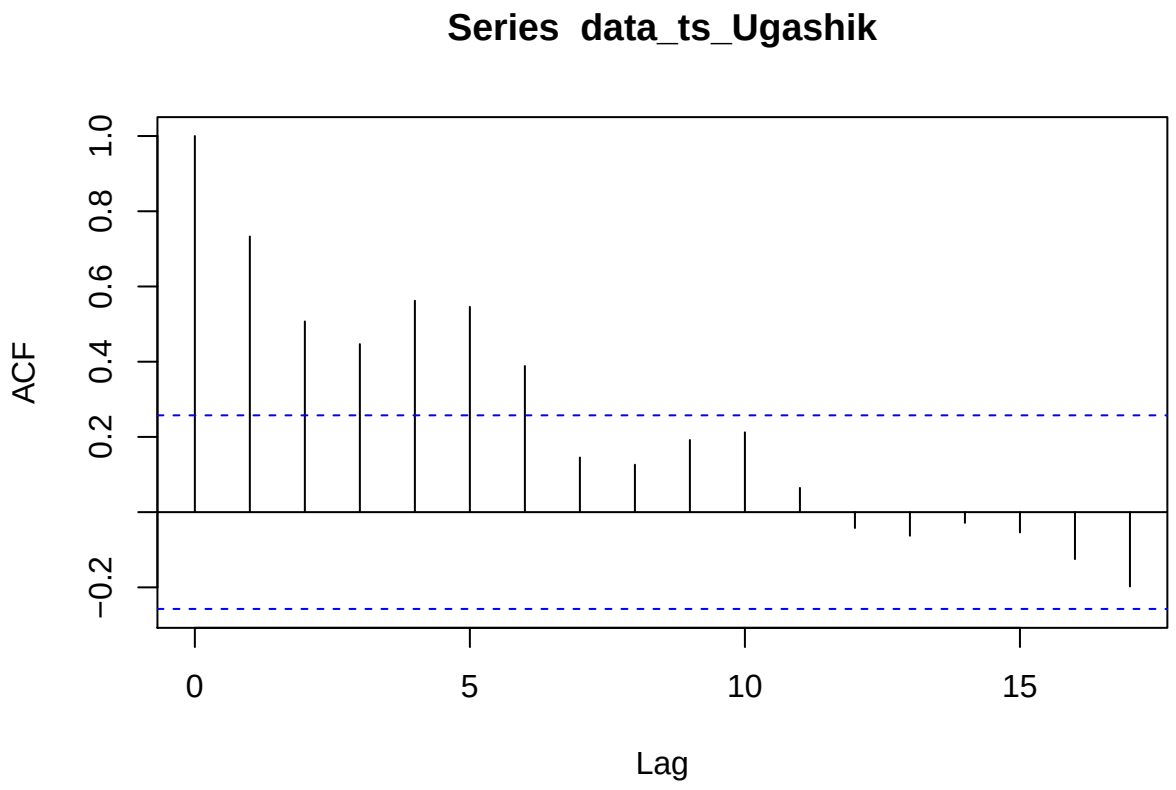


It looks like there is a trend in Wood river log returns and maybe some negative autocorrelation. Variance in Ugashik seems to be higher until 1980 and lower afterwards.

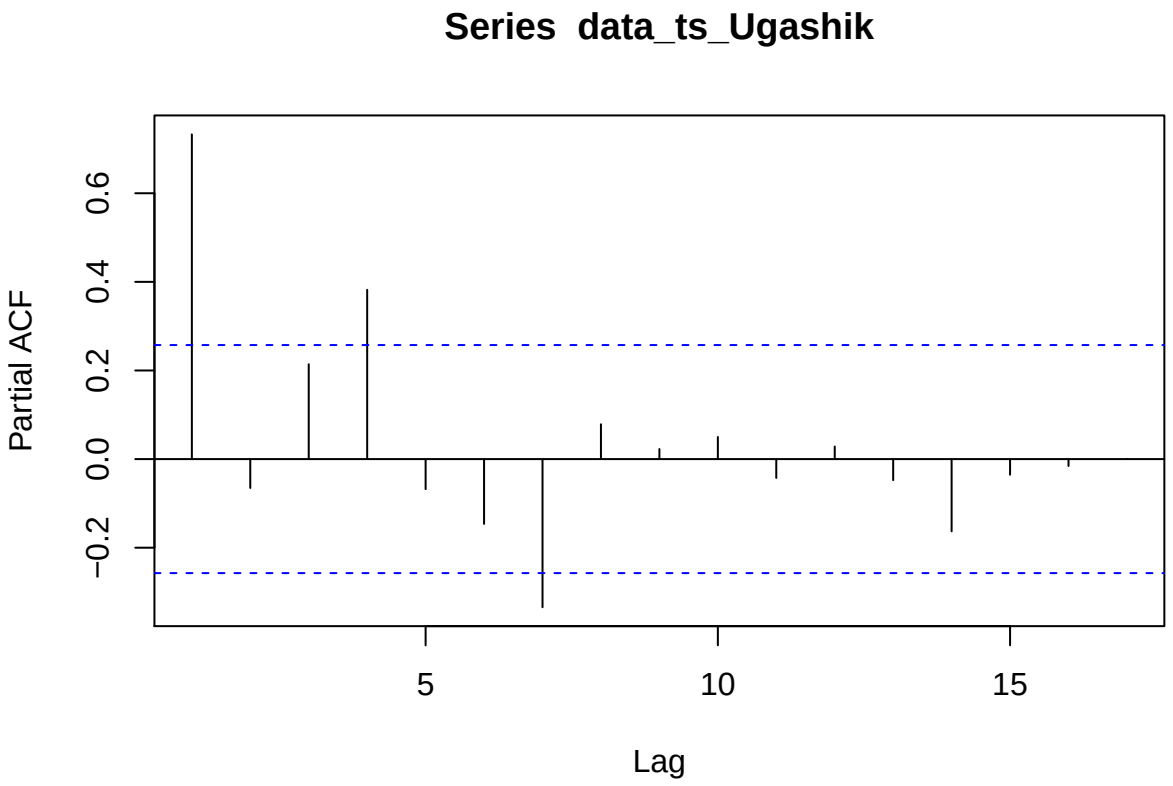
Use ACF and PACF

Use the ACF and PACF plots to explore the time series. What did you learn? Also try decomposition.

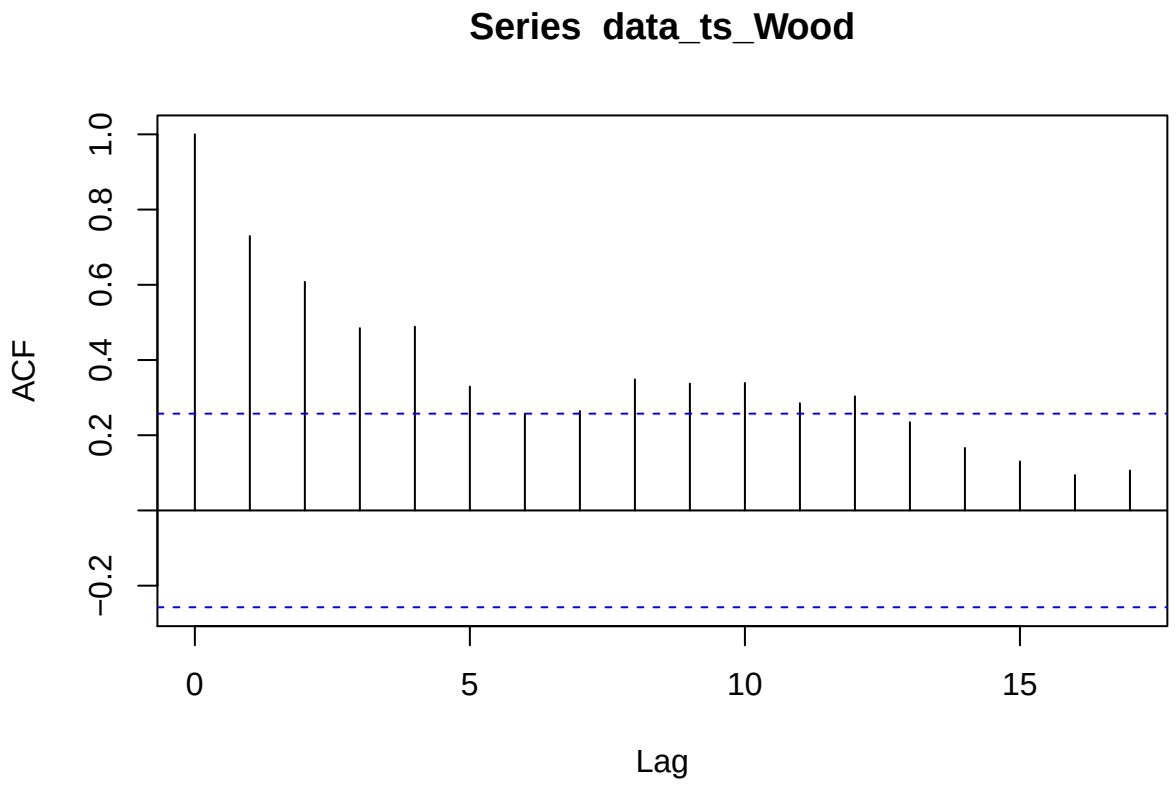
```
data_ts_Ugashik <- ts(subdata$lntotal[subdata$system == "Ugashik"],
                      start=subdata$ret_yr[subdata$system == "Ugashik"][1])
data_ts_Wood <- ts(subdata$lntotal[subdata$system == "Wood"],
                      start=subdata$ret_yr[subdata$system == "Wood"][1])
train_Ugashik <- window(data_ts_Ugashik, 1963, 2005)
test_Ugashik <- window(data_ts_Ugashik, 2006, 2020)

train_Wood <- window(data_ts_Wood, 1963, 2005)
test_Wood <- window(data_ts_Wood, 2006, 2020)

acf(data_ts_Ugashik)
```



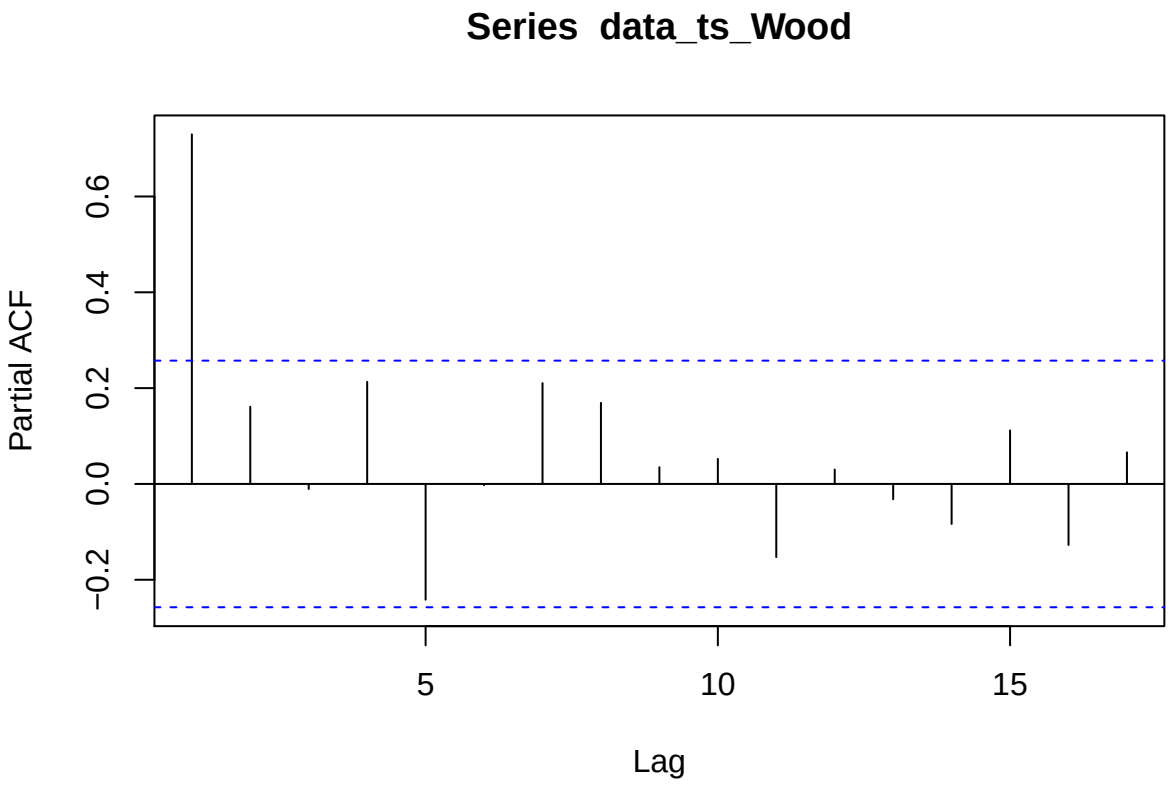

```
pacf(data_ts_Ugashik)
```




```
acf(data_ts_Wood)
```




```
pacf(data_ts_Wood)
```



The acf plots for both rivers are both slowly decaying and the pacf plots for both rivers show significants at various lags. In the Ugashik river lags 1, 4 and 7 are significant whereas in the Wood river, only lag 1 is significant. There might be AR1 structure in the ata from both rivers with some potentially higher order AR terms in Ugashik.

Test for stationarity

Run tests and discuss any stationarity issues and how these were addressed.

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.0.5
```

```
mod_Ugashik <- auto.arima(data_ts_Ugashik)
mod_Ugashik
```

```
## Series: data_ts_Ugashik
## ARIMA(0,1,0)
##
## sigma^2 = 0.7894:  log likelihood = -74.14
## AIC=150.28   AICc=150.35   BIC=152.32
```

```
mod_Wood <- auto.arima(data_ts_Wood)
mod_Wood
```

```
## Series: data_ts_Wood
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.3969  -0.8184
## s.e.  0.1755   0.1041
##
## sigma^2 = 0.218:  log likelihood = -36.7
## AIC=79.4   AICc=79.85   BIC=85.53
```

auto.arima suggests that ARIMA(0,1,0) fits Ugashik river data the best and ARIMA(1,1,1) fits the Wood river data the best.

```
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 4.0.5
```

```
adf.test(data_ts_Ugashik, k = 0)
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_ts_Ugashik
## Dickey-Fuller = -3.5726, Lag order = 0, p-value = 0.0433
## alternative hypothesis: stationary
```

```
kpss.test(data_ts_Ugashik, null =c("Level"))
```

```
## Warning in kpss.test(data_ts_Ugashik, null = c("Level")): p-value smaller than
## printed p-value
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  data_ts_Ugashik
## KPSS Level = 0.79781, Truncation lag parameter = 3, p-value = 0.01
```

```
##Tests give different results
## Ugashik data is probably NON stationary


adf.test(data_ts_Wood, k = 0)
```

```
## Warning in adf.test(data_ts_Wood, k = 0): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  data_ts_Wood
## Dickey-Fuller = -5.0089, Lag order = 0, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(data_ts_Wood, null =c("Trend"))
```

```
## Warning in kpss.test(data_ts_Wood, null = c("Trend")): p-value greater than
## printed p-value
```

```
##
##  KPSS Test for Trend Stationarity
##
## data:  data_ts_Wood
## KPSS Trend = 0.040572, Truncation lag parameter = 3, p-value = 0.1
```

```
#note that this gives different results depending on what null is

#wood is stationary around trend
```

Wood river data is probably stationary around a trend, but tests give differnet results for Ugashik data. Ugashik data is probably non stationary.

```
ndiffs(data_ts_Ugashik, test='kpss')
```

```
## [1] 1
```

```
ndiffs(data_ts_Ugashik, test='adf')
```

```
## [1] 1
```

```
ndiffs(data_ts_Wood, test='kpss')
```

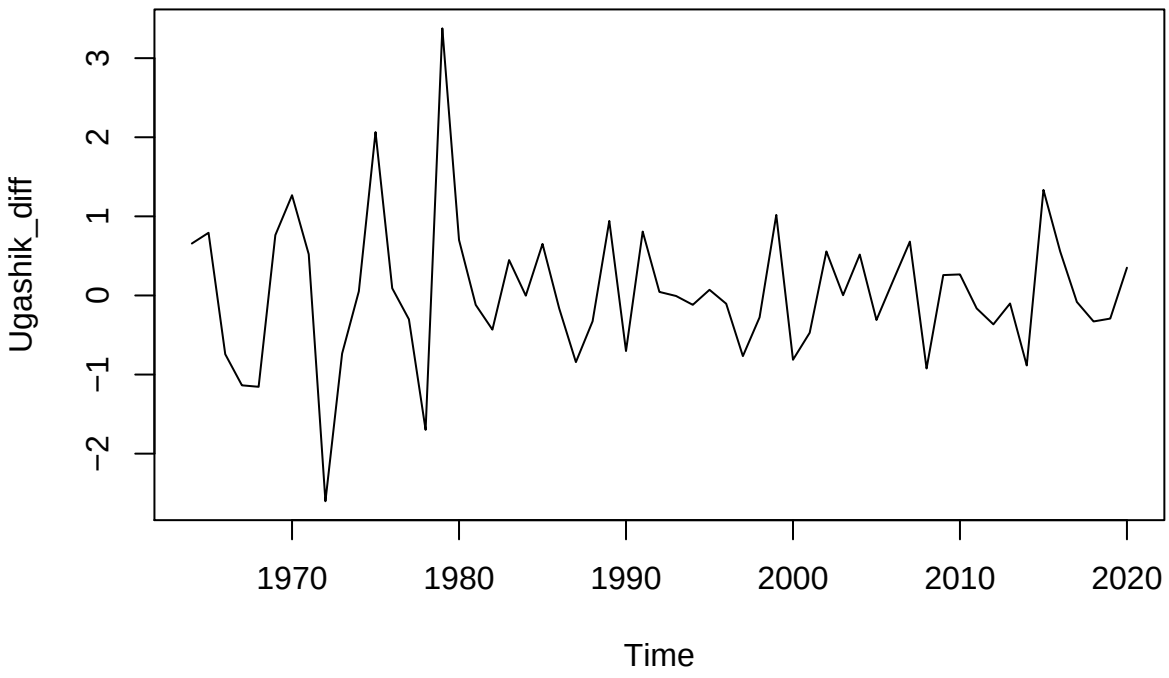
```
## [1] 1
```

```
ndiffs(data_ts_Wood, test='adf')
```

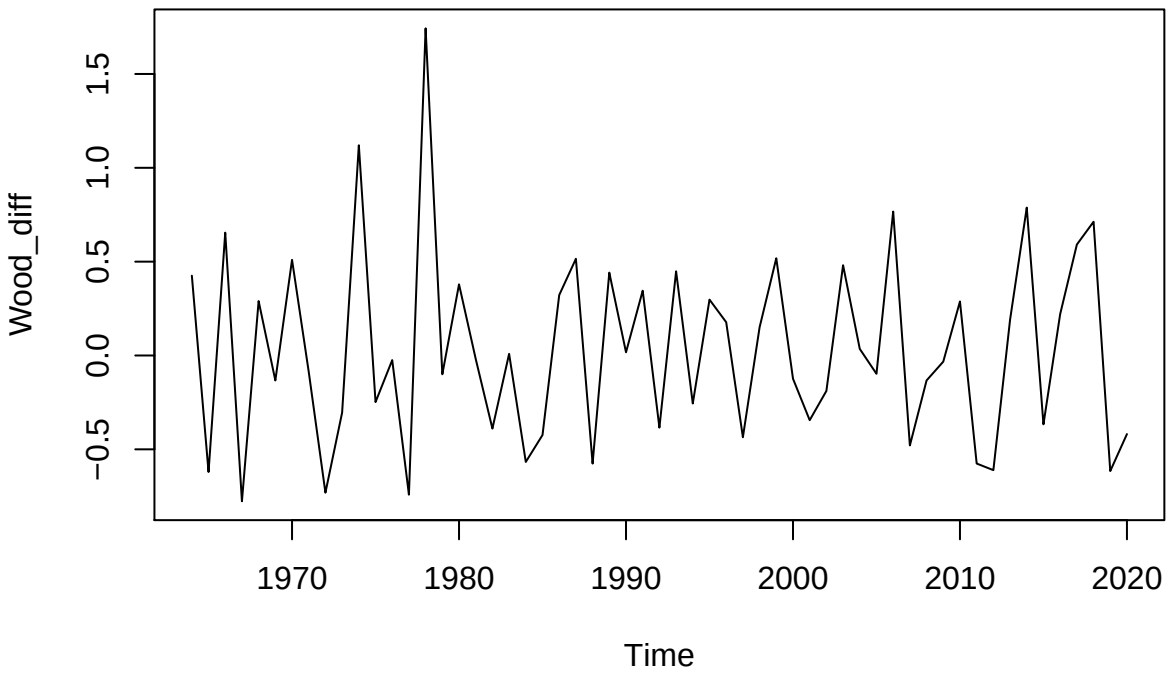
```
## [1] 1
```

Differencing the data

```
Ugashik_diff <- diff(data_ts_Ugashik)
Wood_diff <- diff(data_ts_Wood)

plot.ts(Ugashik_diff)
```




```
plot.ts(Wood_diff)
```



This looks pretty stationary.

Stationarity on differenced data

```
adf.test(Ugashik_diff, k = 0)
```

```
## Warning in adf.test(Ugashik_diff, k = 0): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  Ugashik_diff
## Dickey-Fuller = -7.7412, Lag order = 0, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(Ugashik_diff, null =c("Level"))
```

```
## Warning in kpss.test(Ugashik_diff, null = c("Level")): p-value greater than
## printed p-value

##
##  KPSS Test for Level Stationarity
##
## data:  Ugashik_diff
## KPSS Level = 0.056441, Truncation lag parameter = 3, p-value = 0.1
```

These results say Ugashik data is probably stationary.

```
adf.test(Wood_diff, k = 0)
```

```
## Warning in adf.test(Wood_diff, k = 0): p-value smaller than printed p-value

##
##  Augmented Dickey-Fuller Test
##
## data:  Wood_diff
## Dickey-Fuller = -9.9733, Lag order = 0, p-value = 0.01
## alternative hypothesis: stationary
```

```
kpss.test(Wood_diff, null =c("Level"))
```

```
## Warning in kpss.test(Wood_diff, null = c("Level")): p-value greater than printed
## p-value

##
##  KPSS Test for Level Stationarity
##
## data:  Wood_diff
## KPSS Level = 0.037388, Truncation lag parameter = 3, p-value = 0.1
```

These results say that Wood river data is also probably stationary.

Using auto.arima

```
library(forecast)
mod_Ugashik <- auto.arima(data_ts_Ugashik)
mod_Ugashik
```

```
## Series: data_ts_Ugashik
## ARIMA(0,1,0)
##
## sigma^2 = 0.7894:  log likelihood = -74.14
## AIC=150.28   AICc=150.35   BIC=152.32
```

```
mod_Wood <- auto.arima(data_ts_Wood)
mod_Wood
```

```
## Series: data_ts_Wood
## ARIMA(1,1,1)
##
## Coefficients:
##          ar1      ma1
##       0.3969  -0.8184
## s.e.  0.1755   0.1041
##
## sigma^2 = 0.218:  log likelihood = -36.7
## AIC=79.4   AICc=79.85   BIC=85.53
```

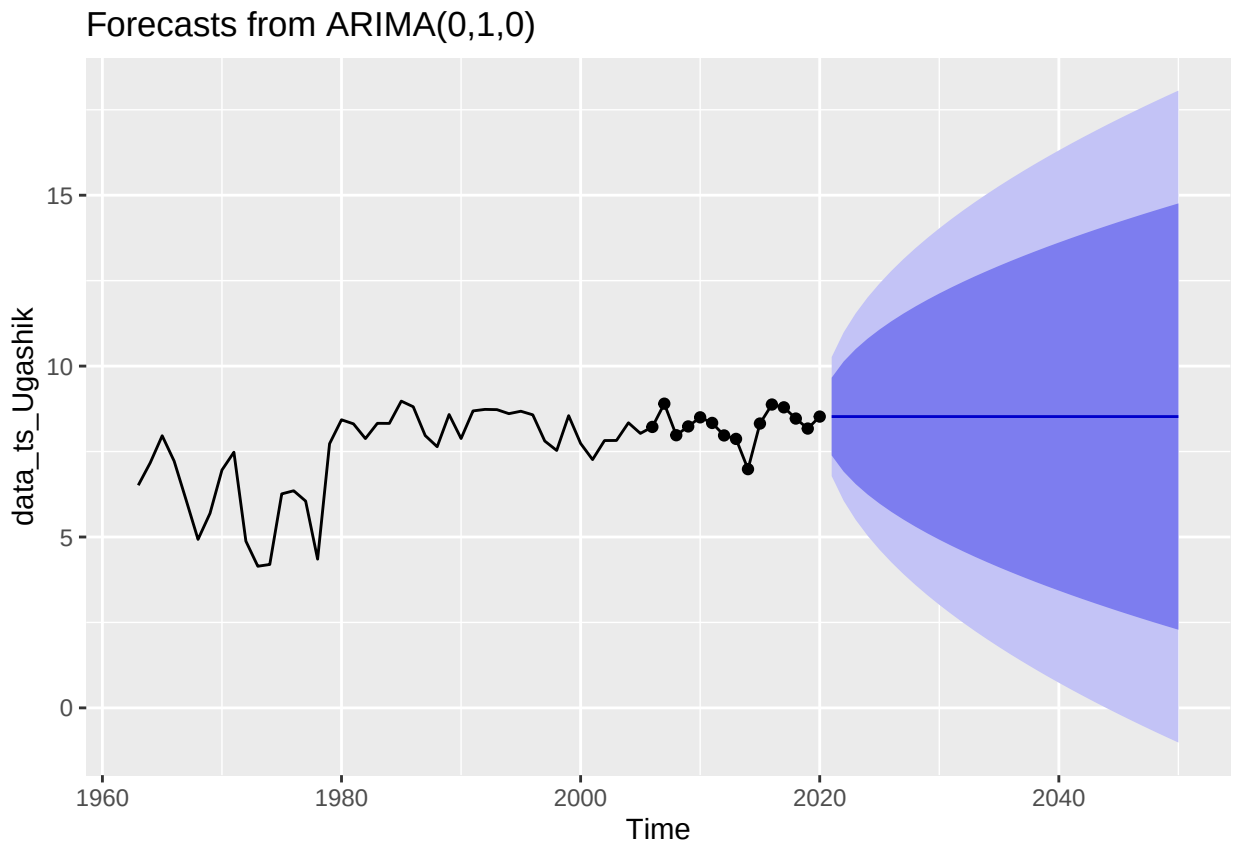
```
library(zoo)
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

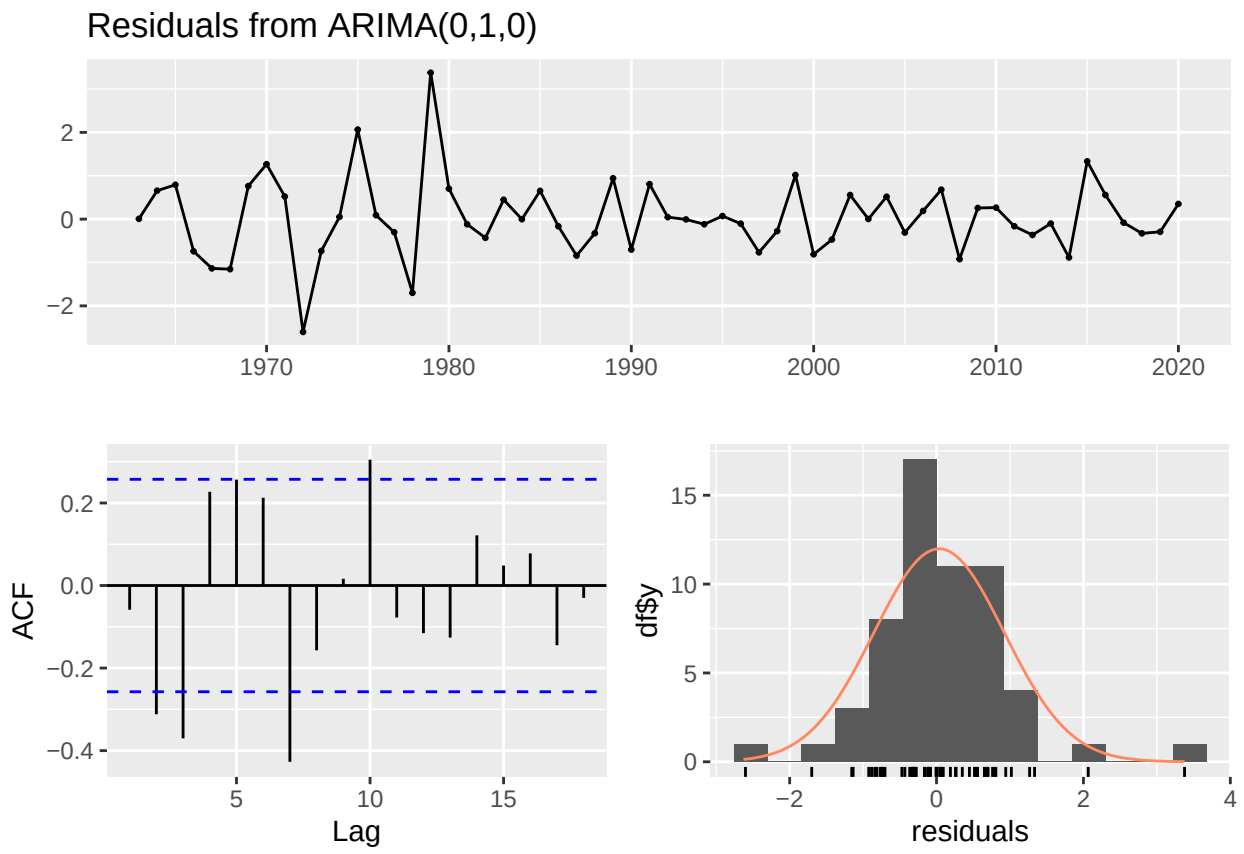
```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
fr <- forecast(mod_Ugashik, h=30)
autoplot(fr) + geom_point(aes(x=x, y=y), data=fortify(test_Ugashik))
```

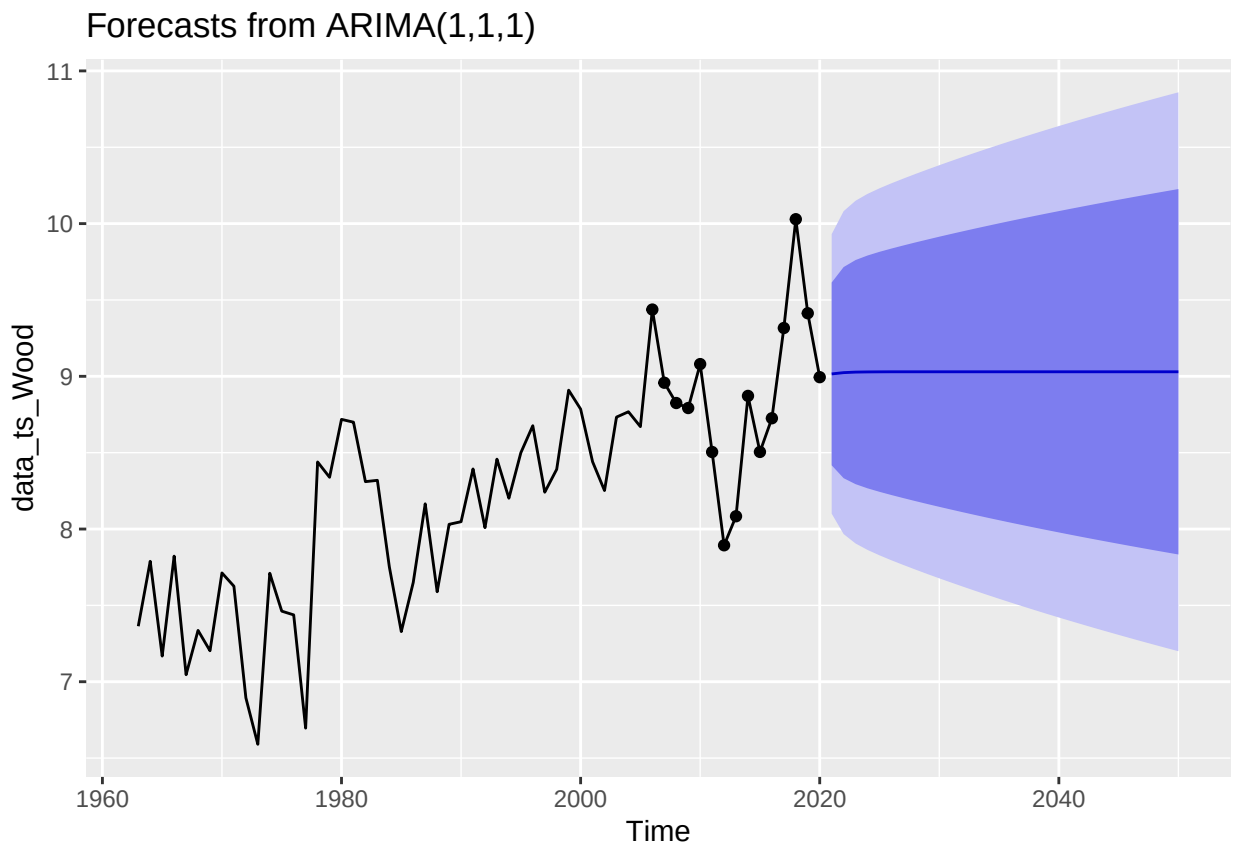



```
checkresiduals(fr)
```

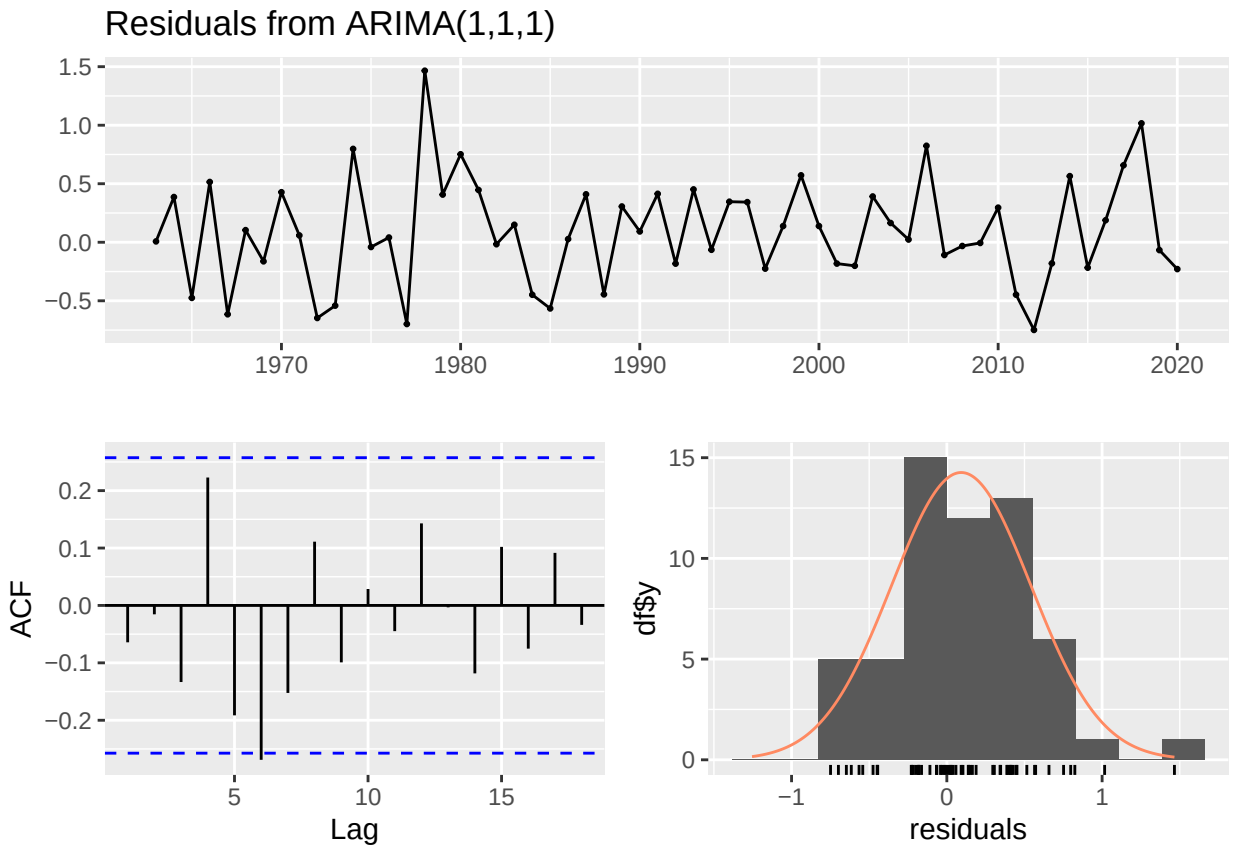



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,0)
## Q* = 46.518, df = 10, p-value = 1.155e-06
##
## Model df: 0.   Total lags used: 10
```

```
library(zoo)
fr_Wood <- forecast(mod_Wood, h=30)
autoplot(fr_Wood) + geom_point(aes(x=x, y=y), data=fortify(test_Wood))
```




```
checkresiduals(fr_Wood)
```

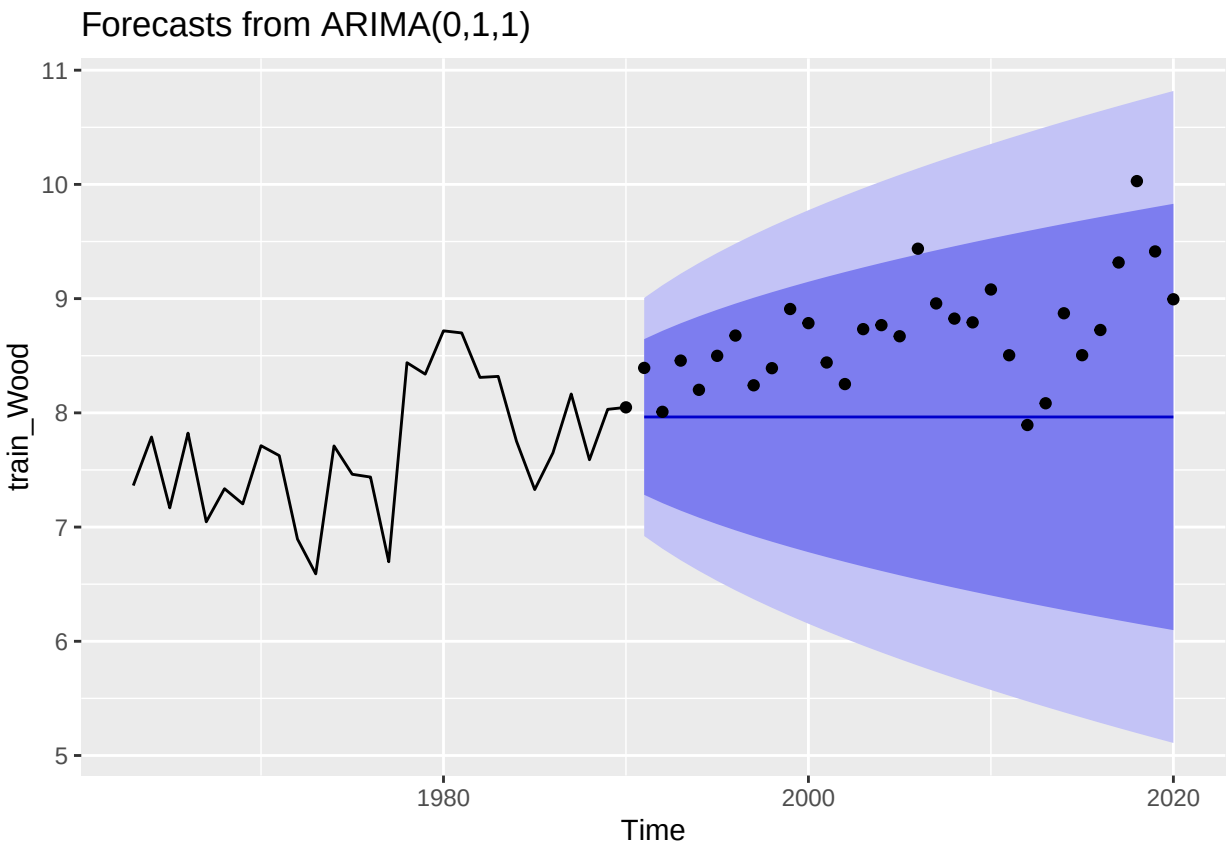



```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,1,1)
## Q* = 15.041, df = 8, p-value = 0.05835
##
## Model df: 2.   Total lags used: 10
```

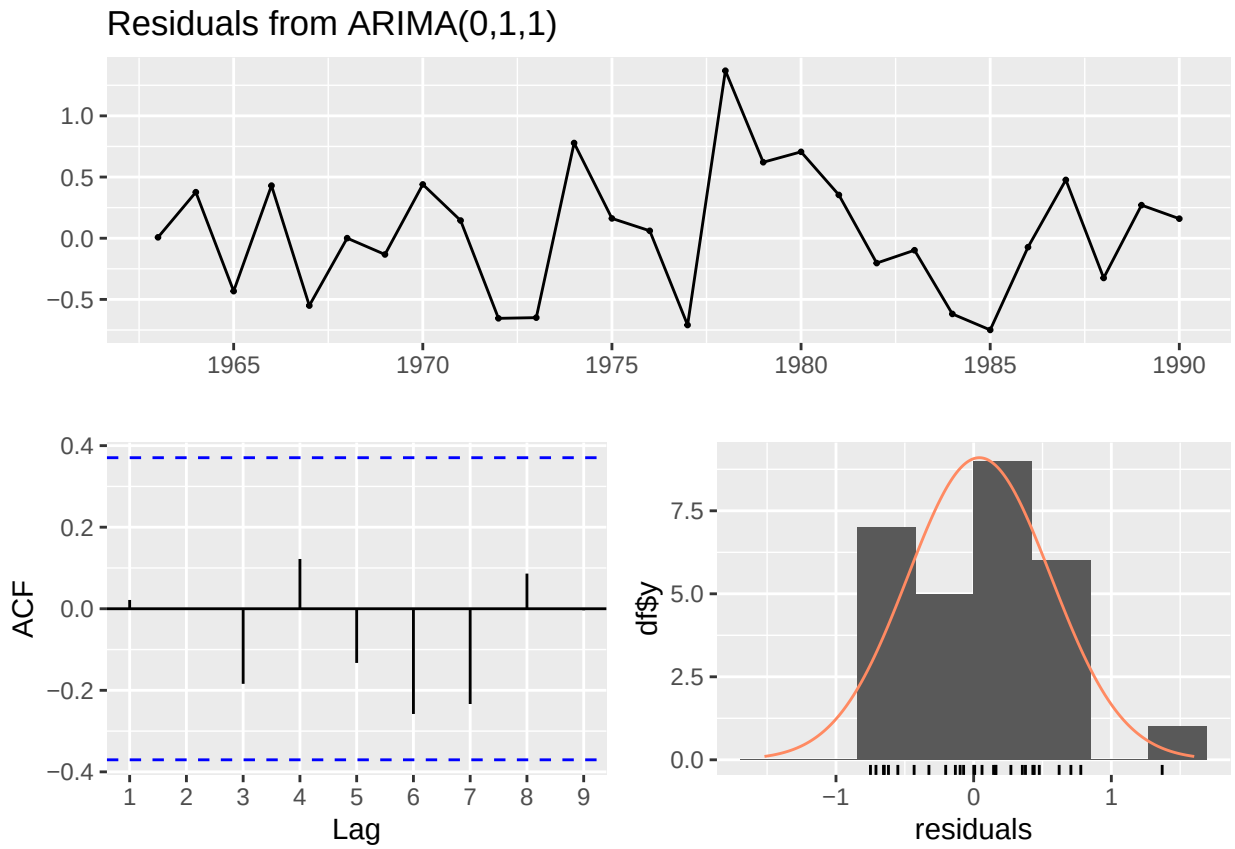
```
library(zoo)
train_Wood <- window(data_ts_Wood, 1963, 1990)
test_Wood <- window(data_ts_Wood, 1990, 2020)
mod_Wood_train <- auto.arima(train_Wood)
mod_Wood_train
```

```
## Series: train_Wood
## ARIMA(0,1,1)
##
## Coefficients:
##           ma1
##       -0.5267
## s.e.   0.2067
##
## sigma^2 = 0.2829:  log likelihood = -20.92
## AIC=45.83   AICc=46.33   BIC=48.43
```

```
fr_Wood_train <- forecast(mod_Wood_train, h=30)
autoplot(fr_Wood_train) + geom_point(aes(x=x, y=y), data=fortify(test_Wood))
```




```
checkresiduals(fr_Wood_train)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)
## Q* = 4.8612, df = 5, p-value = 0.433
##
## Model df: 1.   Total lags used: 6
```

Results

Discussion

Description of each team member's contributions

Example: "All team members helped decide on the goal and ran the analyses for the individual regions. Team members 2 & 3 wrote most of the code for the analysis of the regions. Team member 4 researched approaches for measuring accuracy of forecasts in [Hyndman & Athanasopoulos[OTexts.com/fpp2] and team member 2 added code for that to the methods. Team member 4 also researched tests for stationarity and worked with team member 2 to code that up. Team member 1 worked on the plotting section of the report using and adapting code that team member 3 wrote. All team members helped edit the report and wrote the discussion together."