Buoy - MA615

Yuxi Wang

2020/9/25

My understanding of the question:

First of all, the purpose of this project is to find the evidence of global warming. The source of the data is a single weather buoy in the NOAA National Data Buoy Center. The understanding of this problem lies in finding variables related to global warming based on limited data, and finally modeling to see if it has a rising temperature trend. Finally, we check the correlation between temperature with other factors.

My approach:

At first, I import and merge the 20-year data sets, and tidy the data. Then, find the variables related to global warming in the data set are air temperature and water surface temperature. Finally, two simple methods are used to test whether there is global warming. The first one is to use time series analysis to decompose to get the trend item of temperature. The second one is to use the least squares method to estimate the slope of the data, because the slope is positive means that the change in temperature is rising over time.Finally, we check the correlation between variables, to see if there have other factor that have relationship with the global warming.

```r
#Step1: Import buoy data and tidy data
#In order to read txt file automatically
# make URLs

url1 <- "http://www.ndbc.noaa.gov/view_text_file.php?filename=mlrf1h"
url2 <- ".txt.gz&dir=data/historical/stdmet/"
years <- c(2000:2019)
urls <- str_c(url1, years, url2, sep = "")
filenames <- str_c("mr", years, sep = "")

###  Read the data from the website
N <- length(urls)
colname_1 <- colnames(read.table(urls[1],fill=TRUE, header = TRUE))
colname_2 <- colnames(read.table(urls[7],fill=TRUE, header = TRUE))
for (i in 1:N){
  suppressMessages(  ###  This stops the annoying messages on your screen.  Do this last.
    assign(filenames[i], read.table(urls[i],fill=TRUE ,header = TRUE))
  )
  file <- get(filenames[i])
  if(ncol(file)==17){
```

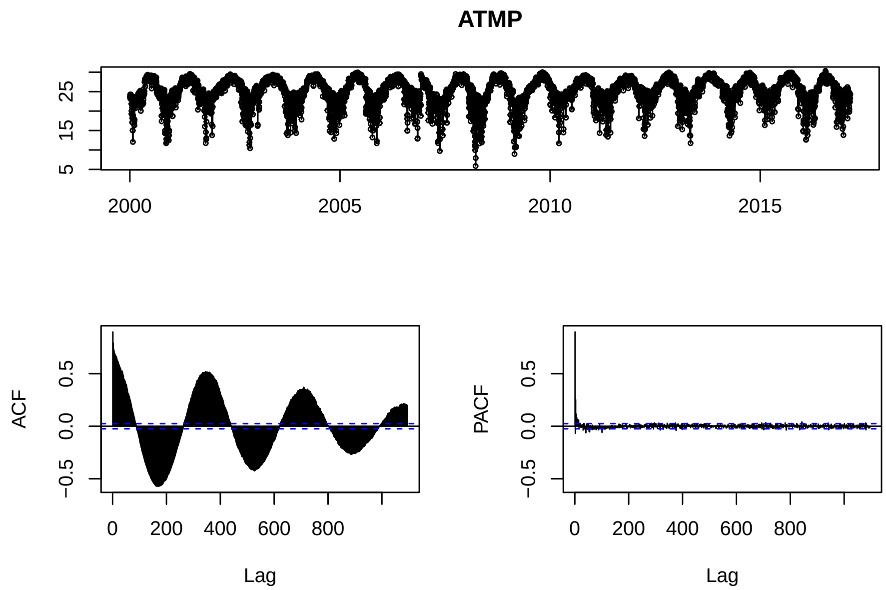
```
    colnames(file)=colname_1
  }else{
    colnames(file)=colname_2
  }
  if(i == 1){
    MR <- file
  }else{
    MR <- dplyr::bind_rows(MR, file)
  }
}
```

This is the first step. In this step, we load the data from the internet and load some useful packages to help me finish the work.

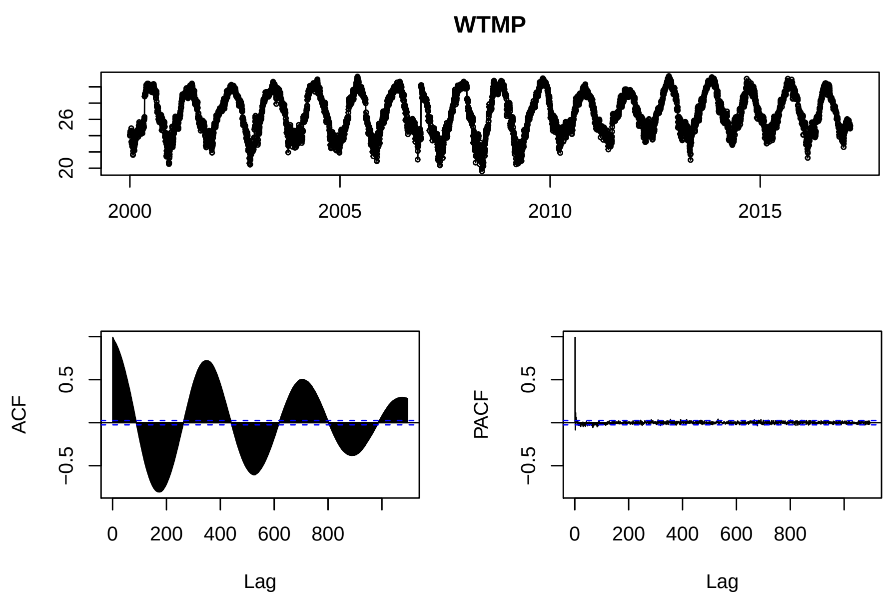
```
# To adjust the date, and deal with outliers in order to get tidy data.
MR <- data.frame(MR)
time <- MR %>% select(YYYY,MM,DD,hh)
time <- make_datetime(MR$YYYY,time$MM,time$DD,time$hh)
time <- data.frame(time)
MR1 <- dplyr::mutate(time,MR)
MR1 <- MR1[,c(-2,-3,-4,-5)]
MR1$time = as.POSIXct(MR1$time)
MR1 <- MR1[,-c(5,6,7,8,12,13,14,15)]
# After watching the output we found that:WVHT,DPD,APD,MWD,DEWP,VIS TIDE,MM are all useless.
MR1 <- filter(MR1,MR$WD <999 &MR$WSPD<99 &MR$GST<99&MR $BAR<9999&MR $ATMP<999&MR $WTMP<999)
summary(MR1)
##       time                          WD             WSPD
##  Min.   :2000-01-01 00:00:00   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:2004-04-18 11:45:00   1st Qu.: 76.0   1st Qu.: 3.700
##  Median :2009-12-13 18:30:00   Median :114.0   Median : 5.800
##  Mean   :2009-04-30 09:57:06   Mean   :137.8   Mean   : 5.974
##  3rd Qu.:2013-12-18 13:15:00   3rd Qu.:183.0   3rd Qu.: 8.100
##  Max.   :2019-03-31 23:00:00   Max.   :360.0   Max.   :33.500
##       GST              BAR              ATMP            WTMP
##  Min.   : 0.000   Min.   : 982.3   Min.   : 3.20   Min.   :15.90
##  1st Qu.: 4.200   1st Qu.:1014.7   1st Qu.:23.40   1st Qu.:24.60
##  Median : 6.500   Median :1016.9   Median :25.60   Median :26.50
##  Mean   : 6.715   Mean   :1016.9   Mean   :25.06   Mean   :26.64
##  3rd Qu.: 9.000   3rd Qu.:1019.0   3rd Qu.:28.00   3rd Qu.:28.90
##  Max.   :41.000   Max.   :1032.4   Max.   :32.20   Max.   :32.10
#sampling the data, reduce data size by using the mean value for each day.
buoy <- MR1%>%
  group_by(date(time))%>%
  summarize(mean(WD),mean(WSPD),mean(GST),mean(BAR),mean(ATMP),mean(WTMP))
## `summarise()` ungrouping output (override with `.groups` argument)
```

This is the second step. In this step, I make substitutions for NA data, transform the date-time data into posix numbers using lubridate. By observing the data, I found that many missing data are replaced by 999 or 9999, so these data need to be filtered out. Also, I use the mean of each day's data to reduce the sample size and prevent outliers.

```r
# transfer the data into time series's form
ATMP <- buoy[,6]
WTMP <- buoy[,7]
ATMP <- ts(ATMP,frequency =365,start=c(2000,1))
tsdisplay(ATMP)
```



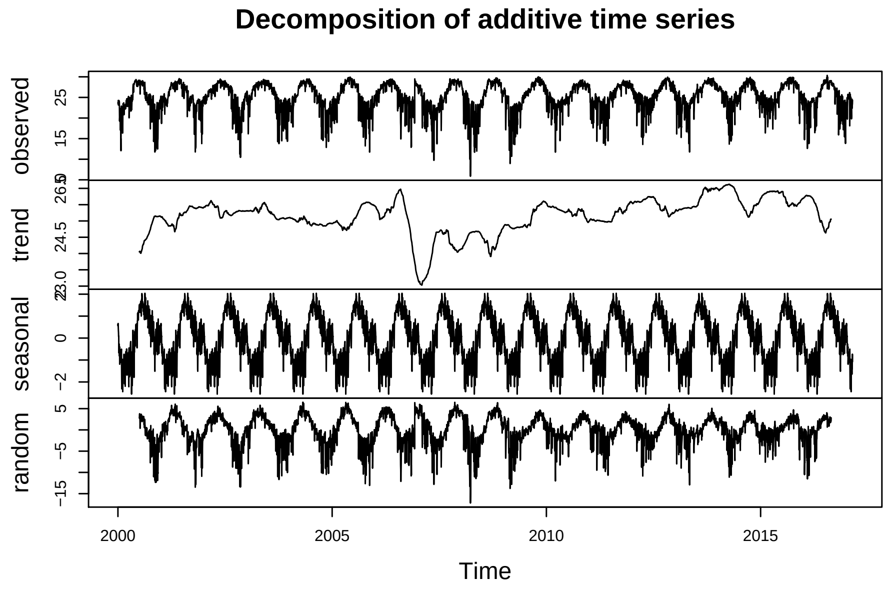

```r
WTMP <- ts(WTMP,frequency =365,start=c(2000,1))
tsdisplay(WTMP)
```



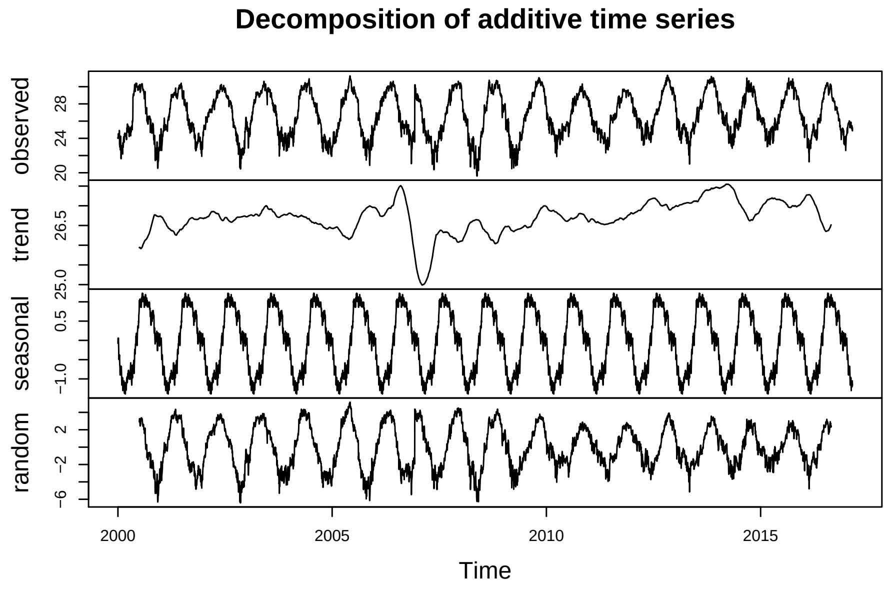
This is step three. In this step, I extract the air temperature and water surface temperature, and transfer them into time series's form. Also, observe their Scatter/ACF/PACF plot.

```
# Using ATMP and WTMP to see whether there has a global warming phenomenon.
#Method one

dc1<-decompose(ATMP)
plot(dc1)
```



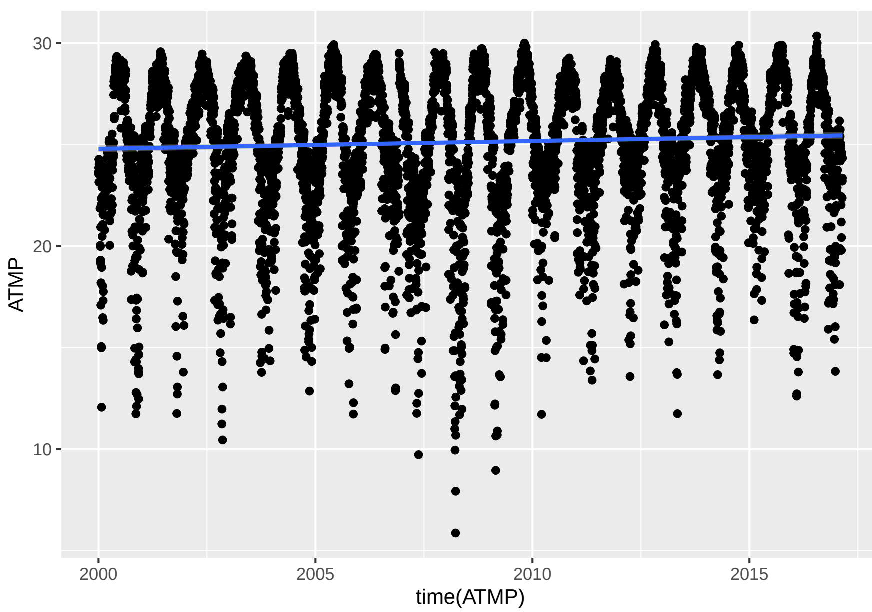

```
dc2<-decompose(WTMP)
plot(dc2)
```



This is step four. In this step, we decompose the ATMP and WTMP. By doing this, we can easily see the trend of every data set.

The conclusion in this step is that, we can see that, after 2018, the air tempreature and water surface tempreature has a rising trend. However, we can not say that our world is warming now. So, we have to do the next part.

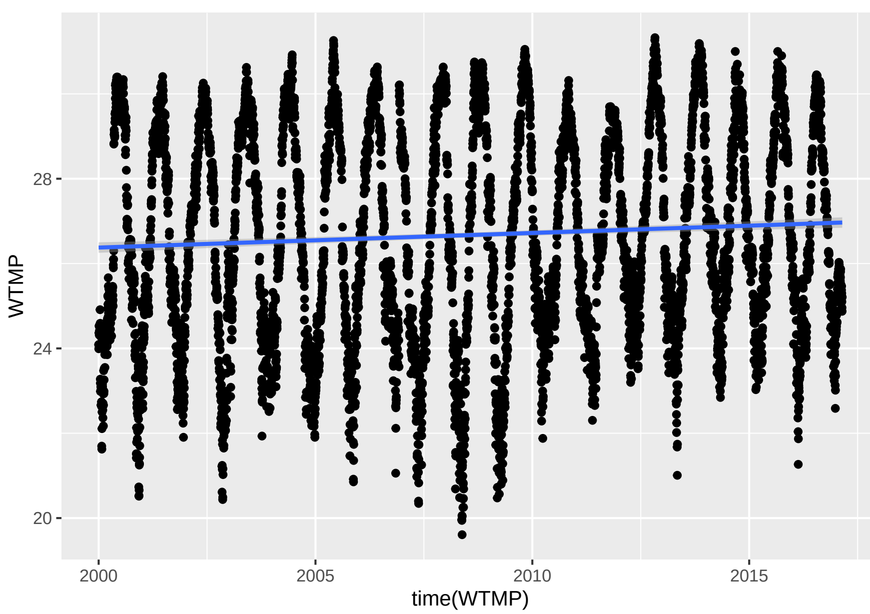
```
#Method 2
model1 <- lm(ATMP~time(ATMP))
summary(model1)
##
## Call:
## lm(formula = ATMP ~ time(ATMP))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.2419  -1.7147   0.4583   2.8675   4.9242
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -52.097885  17.779929  -2.930   0.0034 **
## time(ATMP)    0.038445   0.008852   4.343 1.43e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.467 on 6257 degrees of freedom
## Multiple R-squared:  0.003006,   Adjusted R-squared:  0.002846
## F-statistic: 18.86 on 1 and 6257 DF,  p-value: 1.427e-05
ggplot(data=ATMP, aes(x=time(ATMP), y=ATMP)) +
  geom_point() +
  stat_smooth(method="lm", formula=y ~ x, se=TRUE)
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



This is step five.

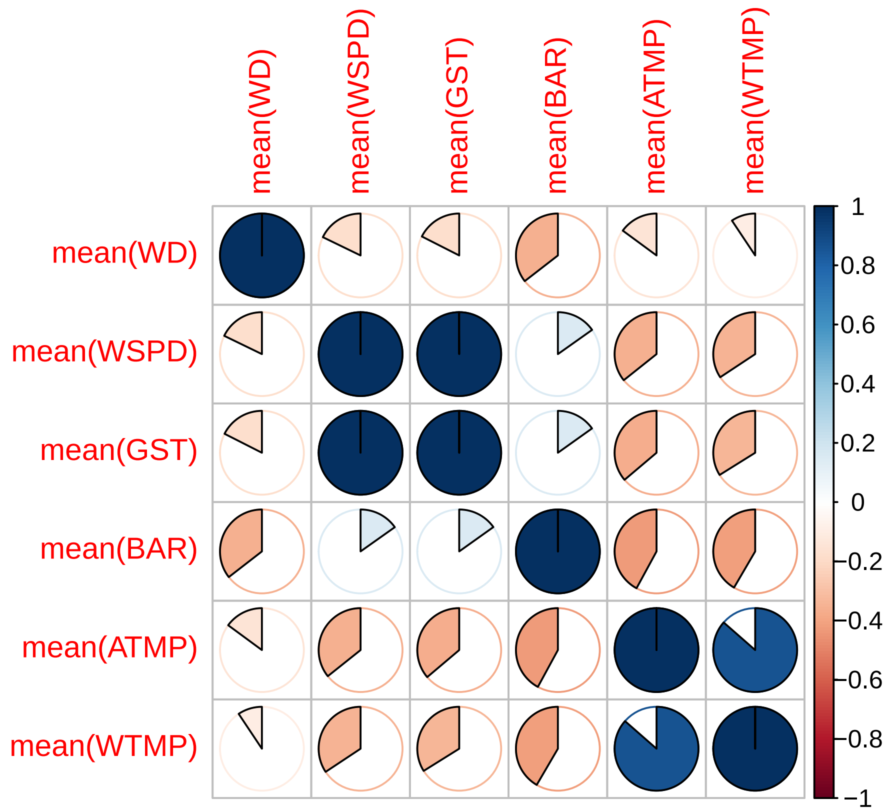
In this plot, we fit the data with ordinary least squares. The estimated coefficience of it is 0.038445. It is clear that the trend of the data is raise. So we can say that the air tempreture is becoming warmer than before.

```
model2 <- lm(WTMP~time(WTMP))
summary(model2)
##
## Call:
## lm(formula = WTMP ~ time(WTMP))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0572 -2.0083 -0.2026  2.3449  4.6948
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.455441  12.796392  -3.318 0.000913 ***
## time(WTMP)    0.034416   0.006371   5.402 6.83e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.495 on 6257 degrees of freedom
## Multiple R-squared:  0.004642,   Adjusted R-squared:  0.004483
## F-statistic: 29.18 on 1 and 6257 DF,  p-value: 6.828e-08
ggplot(data=WTMP, aes(x=time(WTMP), y=WTMP)) +
  geom_point() +
  stat_smooth(method="lm", formula=y ~ x, se=TRUE)
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```



In this plot, we fit the data with ordinary least squares. The estimated coefficience of it is 0.034416. It is clear that the trend of the data is raise. So we can see that the water surface tempreture is becoming warmer than before.

```
buoy1 <- buoy[,-c(1)]
b <- cor(buoy1)
corrplot(b,method="pie")
```



This is step six.

In theory, air temperature and water surface temperature are directly related to global temperature. And we are not sure whether other factors in the data also have a certain degree of correlation with global warming. So I draw a pie chart of the correlation matrix. It can be clearly seen from the figure that there are almost no variables that have a large positive or negative correlation with the two temperature variables. Therefore, it can be roughly considered that the only variables related to global warming in the data of NDBC Station 44013 are air temperature and water surface temperature.

Conclusion:

Our final conclusion is that from the perspective of air temperature and water surface temperature, global warming does exist objectively, although the degree of warming is not drastic. In addition, other data in NDBC Station 44013 are not very relevant to global warming.