Buoy Project

Wendy Liang

2020/9/24

1.Introduction

overview

We need to find evidence of global warming in the data collected by a single weather buoy in the NOAA National Data Buoy Center. In these dataset, there are 18 standard meteorological variables.ATMP, WTMP and DEWP are temperature variable and YY, MM, DD, hh and mm are time variables. In my opinion, what we need to do is finding certain relationship between time and temperature.

approach outline

- step1:import and clean the data
- step2:deal with the time variables
- step3:observe the data by visualization
- step4:explore the data by building regression model
- step5:gain the conclusion

2.Exploratory Data Analysis

I will display how I organize my work in this part.

load all the library

```
library(lubridate)
library(ggplot2)
library(dplyr)
library(xts)
```

import the dataset

The cleaning process of the data is in another R.script.You can find it here link (https://github.com/wendylzw/Buoy-Projet/blob/master/import%20data.R).

```
Buoy=read.csv("Buoydata.csv")
```

Variables

Add two Variables

Firstly,I add a column called *season* in the dataset. (spring = 1, summer = 2, autumn = 3, winter = 4)

Secondly, I Use *Lubridate* package to create a time variable in form “year-month-date-hour”.

```
Buoy$season=0
Buoy$season[Buoy$MM==1|Buoy$MM==2|Buoy$MM==3]=1
Buoy$season[Buoy$MM==4|Buoy$MM==5|Buoy$MM==6]=2
Buoy$season[Buoy$MM==7|Buoy$MM==8|Buoy$MM==9]=3
Buoy$season[Buoy$MM==10|Buoy$MM==11|Buoy$MM==12]=4
Buoy=filter(Buoy,Buoy$hh==12)
time=make_datetime(year = Buoy$X.YY, month = Buoy$MM, day = Buoy$DD)
Buoy=mutate(Buoy,time=time)

#check season
#filter(Buoy,month(time)==5)
```

Describe Variables

Here's the url of the data description link (https://www.ndbc.noaa.gov/measdes.shtml)

ATMP is Air temperature (Celsius).

WTMP is Sea surface temperature (Celsius).

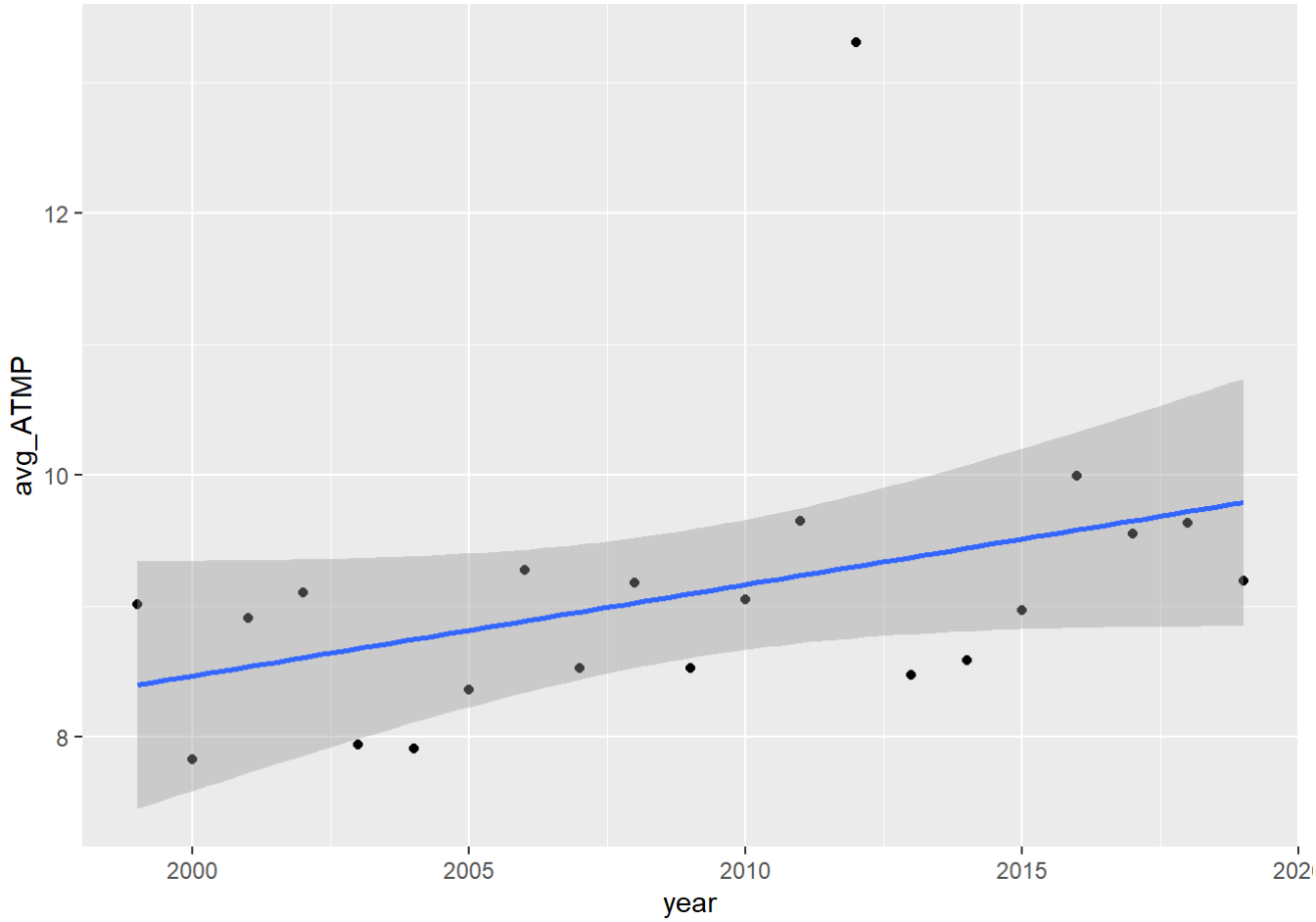
time is the date.

Visulization

year average of ATMP

This is a plot showing the average air temperature of each year. In addition, the smooth line also show the positive correlation between year and the temperature. In other word, the temperature dose increase over time.

```
#year average of ATMP
plane=group_by(Buoy,X.YY)
avg=summarize(plane,avg_ATMP = mean(ATMP))
ggplot(avg,aes(x=X.YY,y=avg_ATMP))+geom_point()+geom_smooth(method ="lm")+xlab("year")
```



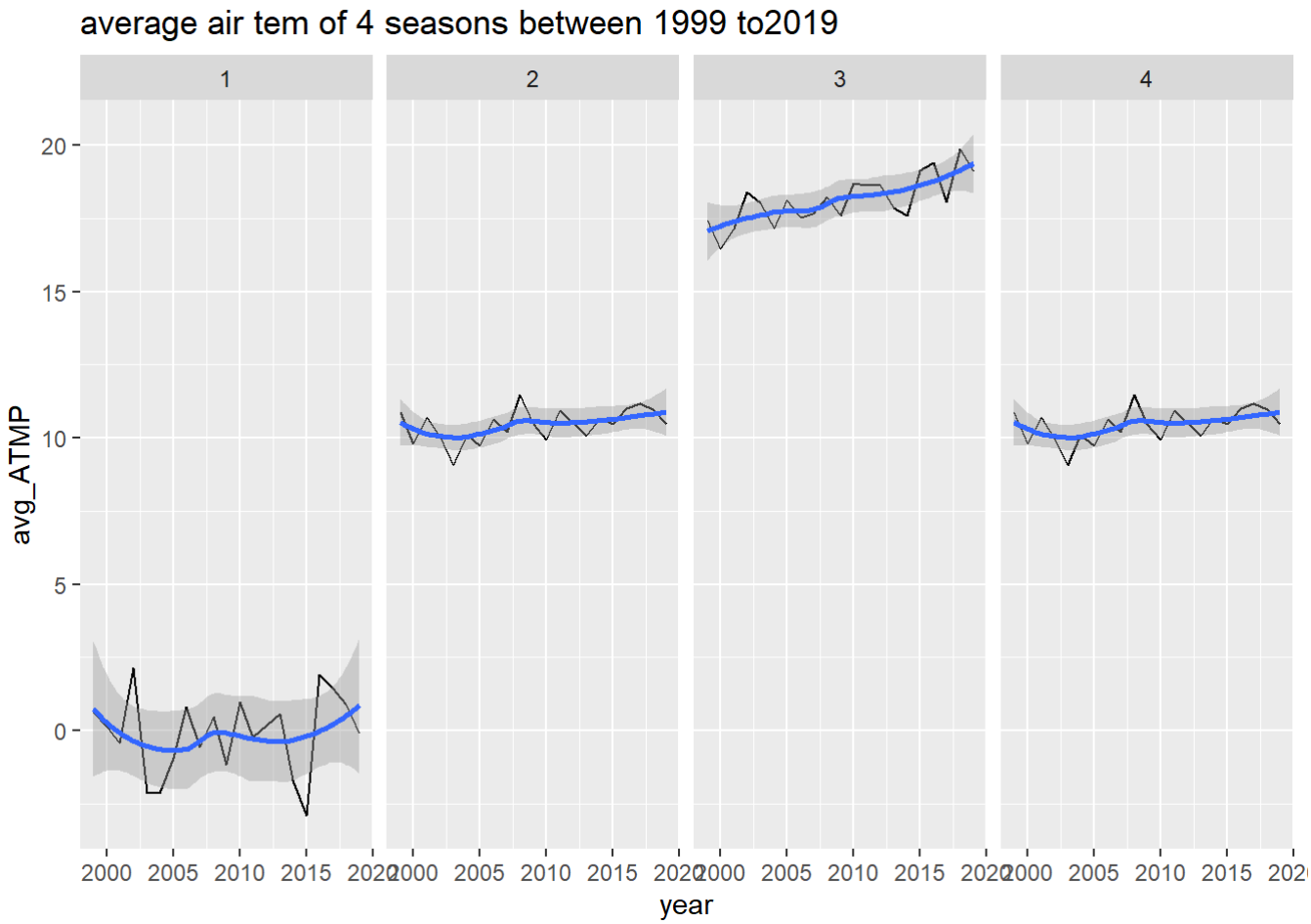
season average of ATMP

This plot shows the changing curve of the average air temperature of each season, from 1999 to 2019. In addition, the smooth line also show the positive correlation between year and the temperature.

```
#sping
sp=filter(Buoy,Buoy$season==1)
plane1=group_by(sp,X.YY)
avg_sp=summarise(plane1,avg_ATMP = mean(ATMP))
avg_sp=mutate(avg_sp,season=1)
#summer
su=filter(Buoy,Buoy$season==2)
plane2=group_by(su,X.YY)
avg_su=summarise(plane2,avg_ATMP = mean(ATMP))
avg_su=mutate(avg_su,season=2)
#autum
aut=filter(Buoy,Buoy$season==3)
plane3=group_by(aut,X.YY)
avg_aut=summarise(plane3,avg_ATMP = mean(ATMP))
avg_aut=mutate(avg_aut,season=3)
#winter
win=filter(Buoy,Buoy$season==4)
plane4=group_by(su,X.YY)
avg_win=summarise(plane4,avg_ATMP = mean(ATMP))
avg_win=mutate(avg_win,season=4)

r=rbind(avg_sp,avg_su,avg_aut,avg_win)

ggplot(r,aes(x=X.YY,y=avg_ATMP))+geom_line()+geom_smooth(methond="lm")+xlab("year")+ggtitle("av
erage air tem of 4 seasons between 1999 to2019")+ facet_grid(. ~ season)
```




```
#if choose special day to represent each season, then...
#spr=filter(Buoy,MM==3,DD==20)
#su=filter(Buoy,MM==6,DD==21)
#aut=filter(Buoy,MM==9,DD==22)
#win=filter(Buoy,MM==12,DD==21)
#Buoy_sea=rbind(spr,su,aut,win)
#plot( xts(spr$ATMP, as.Date(spr$time, format='%Y/%m/%d')),type = 'l',main=' ')
```

In this plot, I find that:

- the temperature in 4 seasons all increases over time
- the temperature in autumn increases most obviously over time
- the temperature in spring increases least obviously over

Regression Model

calculate the time length to now

I don't think time in "ymd" form is suitable for regression model, so I change it to numeric form —- **to_now**.

to_now means the time length from certain date to now.

to_now_st means the standard form of **to_now**

```
now=ymd("2020-09-25")
int=interval(Buoy$time,now)
to_now=time_length(int,unit="day")
Buoy=mutate(Buoy,to_now=time_length(int,unit="day"),to_now_st=(to_now-mean(to_now))/sd(to_now))
```

build different model, based on “year”

```
#everyday model
fit1=lm(ATMP~to_now_st,data=Buoy)
summary(fit1)
```

```
##
## Call:
## lm(formula = ATMP ~ to_now_st, data = Buoy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7684  -5.8250   0.3116   7.1060  15.3526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.99205    0.09257  97.137  < 2e-16 ***
## to_now_st   -0.52671    0.09258  -5.689 1.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.952 on 7378 degrees of freedom
## Multiple R-squared:  0.004368,   Adjusted R-squared:  0.004233
## F-statistic: 32.37 on 1 and 7378 DF,  p-value: 1.323e-08
```

```
#year average model
avg=summarise(group_by(Buoy,X.YY),avg_ATMP = mean(ATMP),avg_now = mean(to_now))
fit2=lm(avg_ATMP~avg_now,data=avg)
summary(fit2)
```

```
##
## Call:
## lm(formula = avg_ATMP ~ avg_now, data = avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9061 -0.6068 -0.1151  0.3956  3.9822
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.8893351  0.4912871  20.129 2.83e-14 ***
## avg_now     -0.0001941  0.0001054  -1.841   0.0813 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 19 degrees of freedom
## Multiple R-squared:  0.1513, Adjusted R-squared:  0.1067
## F-statistic: 3.388 on 1 and 19 DF,  p-value: 0.08134
```

```
#transform:standard year average model
a=(avg$avg_now-mean(avg$avg_now))/sd(avg$avg_now)
fit3=lm(avg_ATMP~a,data=avg)
summary(fit3)
```

```
##
## Call:
## lm(formula = avg_ATMP ~ a, data = avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9061 -0.6068 -0.1151  0.3956  3.9822
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.0937     0.2335  38.942   <2e-16 ***
## a            -0.4405     0.2393  -1.841   0.0813 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.07 on 19 degrees of freedom
## Multiple R-squared:  0.1513, Adjusted R-squared:  0.1067
## F-statistic: 3.388 on 1 and 19 DF,  p-value: 0.08134
```

```
#transform: log year average model
fit4=lm(avg_ATMP~log(avg_now),data=avg)
summary(fit4)
```

```
##
## Call:
## lm(formula = avg_ATMP ~ log(avg_now), data = avg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9172 -0.6786 -0.0962  0.3081  4.1565
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   12.7983     2.5859   4.949  8.9e-05 ***
## log(avg_now)  -0.4574     0.3179  -1.439    0.166
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.103 on 19 degrees of freedom
## Multiple R-squared:  0.09826,    Adjusted R-squared:  0.0508
## F-statistic:  2.07 on 1 and 19 DF,  p-value: 0.1665
```

According to the coefficient and R^2, I think the second model is the best.

We have $ ATMP=-0.44*time+9.09$ and the $R^2 = 0.15$.In fact, this model fits no well. But we can know that the ATMP increases over times since the slope is < 0.

build different model, based on “season”

Since “autumn” is the best fitted season through the visualization, I just choose its data build regression model. We can use the same method on other three seasons analysis.

```
#aut
aut=filter(Buoy,Buoy$season==3)
plane3=group_by(aut,X.YY)
avg_aut=summarise(plane3,avg_ATMP = mean(ATMP),avg_now = mean(to_now))
avg_aut=mutate(avg_aut)

#season average model
fit5=lm(avg_ATMP~avg_now,data=avg_aut)
summary(fit5)
```

```
##
## Call:
## lm(formula = avg_ATMP ~ avg_now, data = avg_aut)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0184 -0.4900  0.2079  0.4107  0.9367
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.922e+01  2.682e-01  71.644  < 2e-16 ***
## avg_now     -2.670e-04  5.804e-05  -4.601 0.000195 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5882 on 19 degrees of freedom
## Multiple R-squared:  0.527,  Adjusted R-squared:  0.5021
## F-statistic: 21.17 on 1 and 19 DF,  p-value: 0.0001951
```

```
#transform: standard season average model
aa=(avg_aut$avg_now-mean(avg_aut$avg_now))/sd(avg_aut$avg_now)
fit6=lm(avg_ATMP~aa,data=avg_aut)
summary(fit6)
```

```
##
## Call:
## lm(formula = avg_ATMP ~ aa, data = avg_aut)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0184 -0.4900  0.2079  0.4107  0.9367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.1341     0.1284 141.272  < 2e-16 ***
## aa           -0.6052     0.1315  -4.601 0.000195 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5882 on 19 degrees of freedom
## Multiple R-squared:  0.527,  Adjusted R-squared:  0.5021
## F-statistic: 21.17 on 1 and 19 DF,  p-value: 0.0001951
```

```
#transform: log season average model
fit7=lm(avg_ATMP~log(avg_now),data=avg_aut)
summary(fit7)
```

```
##
## Call:
## lm(formula = avg_ATMP ~ log(avg_now), data = avg_aut)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0610 -0.4414 -0.0740  0.5259  0.7834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   24.1039     1.3814  17.449 3.75e-13 ***
## log(avg_now)  -0.7388     0.1702  -4.341 0.000352 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.606 on 19 degrees of freedom
## Multiple R-squared:  0.498,  Adjusted R-squared:  0.4716
## F-statistic: 18.85 on 1 and 19 DF,  p-value: 0.0003516
```

The R^2 of the first two model are both better than the third one, so we can use them to describe the data.(We don't need to use log.)0

We have $ATMP = -0.61 * time + 18.1$ and the $R^2 = 0.53$. This "season model" fits much better than "year model". There are stronger positive correlation between time and temperature. We can also know that the autumn ATMP increases over time.

By the way, We can use the same method to analyze other three seasons.I plan to not display in this R.script.

3.Conclusions

We can find evidence of global warming from these dataset.