El-Nino dataset report

1. Introduction:

ENSO is the change in water temperature in the Pacific Ocean, and it is three types: El Nino which is a term to describe warming of water temperature, El Nina for cooling water temperature, and a neutral temperature. It caused many problems for many countries. The cycle of 1982-1986 was the strangest in the century. For regions in the west of the pacific, they expressed fires, while USA and Peru expressed floods and increasing in rainfall.

TAO array is an ocean observation system that was used to study the interactions in the ocean atmosphere. It allows scientists to have weather predictions, by forecast the weather basing on moored buoys, drifting buoys, volunteer ship temperature probes, sea level measurements. And this would help scientists to predict when the next cycle on ENSO will happened.

The aim of this report is to explore El-nino data set and investigate relationships between variables using data analytics methods learned in this module.

2. Dataset description:

This database contains 2 files: the first one contain the columns names, the second one for our data. It have 12 column and 178080 row. The following table is a summary on this dataset:

obs	Observation, integer from 1 to 178080
year	Years, integer from 80 to 98
month	Months, integer from 1 to 12
day	Days, integer from 1 to 31
date	Dates, integer from 800307 to 980623
latitude	Latitude, numeric from -8.81 to 9.05
longitude	longitude, numeric from -180 to 171.08
zon.winds	Zonal winds, numeric from -12.4 to 14.3 (west<0, east>0), have 25163 missing data
mer.winds	Meridian winds, numeric from -11.60 to 13 (south<0, north>0), have 25162 missing data
humidity	Humidity, numeric from 45.4 to 99.9, have 65761 missing data
air temp.	Air temperature, numeric from 17.05 to 31.66, have 18237 missing data
s.s.temp.	Sea surface temperature, numeric from 17.35 to 31.26, have 17007 missing data

The missing data in this database was represented as "."

3. Data cleaning:

The cleaning started by giving names to the columns and replacing the missing data "." by "NA", then transform the date format from integer to date, also transform the format of longitude. After that create new variable: hemisphere: if latitude is less than 0 that means it's south, else it's north.

After that I have made a subset data frame: I deleted the column of humidity because it have a large amount of missing data (37% of it is missing), and deleted rows of missing data in air temperature and sea surface temperature. The subset data now have 148956 row and 13 column.

4. Data exploration:

4.1. Correlation:

In order to discover relations between variables, correlation plot was used as it is shown in figure 1. The plot shows a strong positive correlation between air temperature and sea surface temperature with 0.95. And between longitude and sea surface temperature, sea surface temperature and meridian with -0.66, -0.68 and 0.40 respectively. While there was weak correlation between zonal and meridian winds with both air temperature and sea surface temperature.

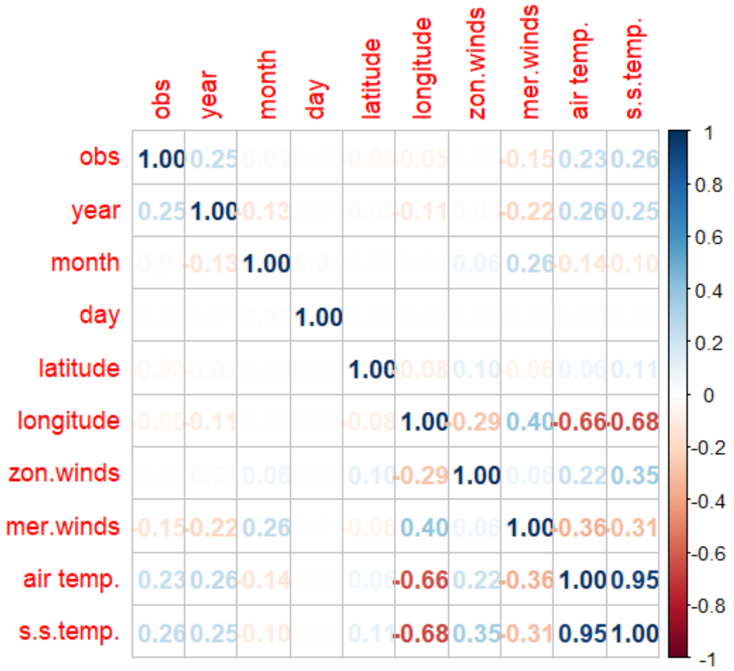

Figure 1: correlation plot

4.2. Scatter plots:

To better visualize relations between correlated variables, the plots in figure 2 were produced. The plot 1 confirms that there is high relation between sea surface temperature and air temperature: when air temperature goes higher the sea surface temperature goes also higher. The plot shows also that there is a relation between latitude and the increase of temperature: the sea surface temperature of the north zones at the same air temperature is higher than the temperature in the south.

Plots 2 and 3 show that the temperature in both air and sea surface have small and high interval of temperature in the west, between 31 and 25 degree. When advancing toward the east the interval of temperature starts getting larger, till it arrive to interval of around [17,31]. And I can say that the average of the west pacific is hotter, and gets less hot when moving toward the east.

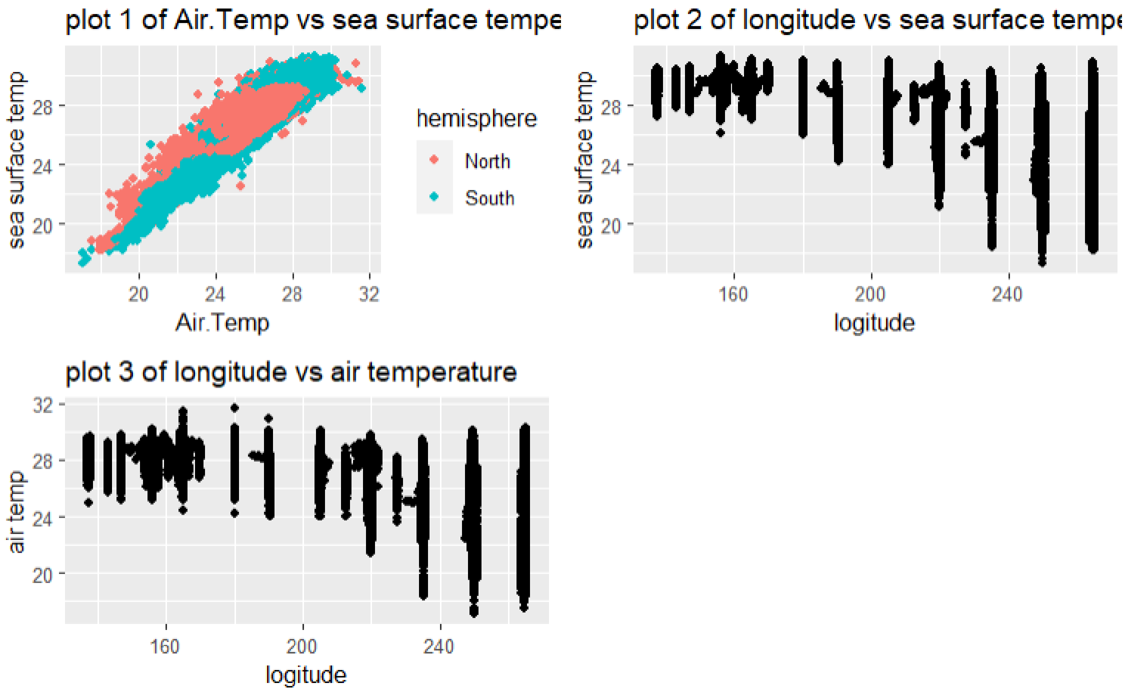

Figure 2: relation between temperature and position

4.3. Time series:

To get the time series plot, I decided to get it for the year of 1994 because in this year where there was the highest number of observations. Figure 3 shows the plot of time series. The plots are showing that the temperature for air and sea surface are having the same trends such as dropping in the temperature in the 8th month, while there is some fluctuations in Humidity and zonal winds.

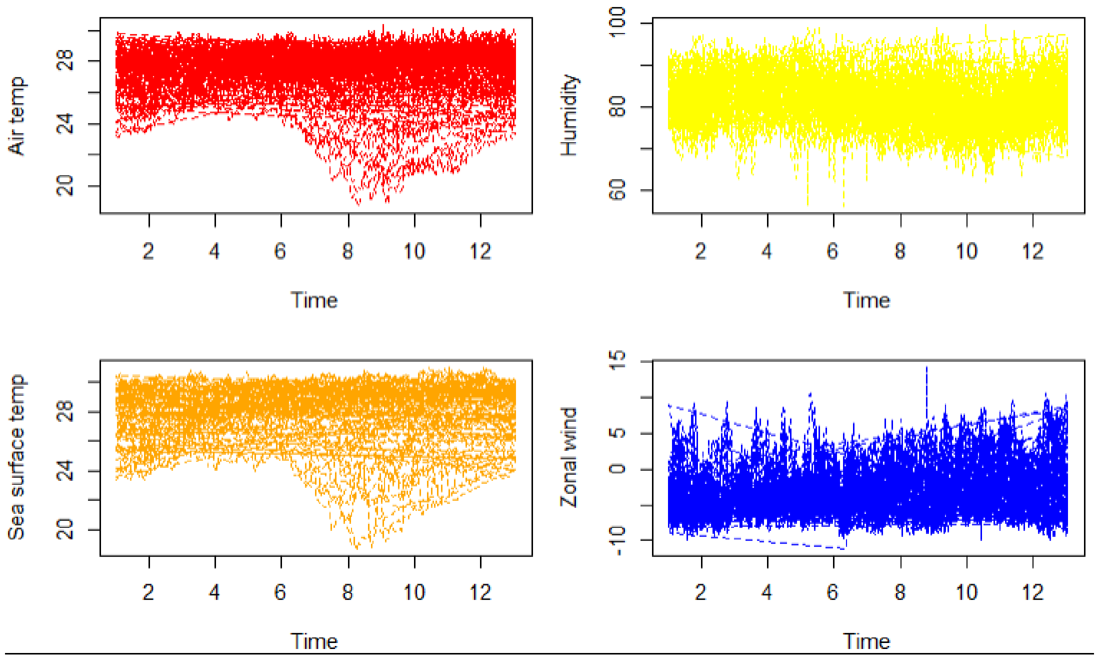

Figure 3: time series

5. Conclusion:

In conclusion, temperature of the air and sea surface are related, when one of them goes higher, the other one will also go high. The position is also related to the change of temperature; the sea surface temperature in the north (latitude>0) goes high faster than in the south. The west of the pacific ocean have smaller interval of temperature that gets bigger when moving towards the east, in general the west is hotter than the east.

The challenge of this dataset was the huge number of missing data (40% of the values are NAs).

References:

“DISASTERS EXPLAINED: EL NIÑO, Discover more about El Niño and how it can impact weather patterns around the globe“ ShelterBox available at: https://shelterbox.org/disasters-explained/el-nino/?gclid=Cj0KCQiAnfmsBhDfARIsAM7MKi0F7Xv0ufSo6SE0ji0ttDsD1J8L76UCwXon0_u5FlAAyiXxEDX6IlgaAlOXEALw_wcB

Example code:

```r
library(ggplot2)
library(GGally)
library(gridExtra)
library(reshape)
library(corrplot)
data_col= read.table('tao-all2.col', sep=",")# read the files
nino= read.table('tao-all2.dat', header= FALSE,na.strings = ".")
colnames(nino)=data_col$V1 # name the columns
str(nino);summary(nino)
nino$longitude= nino$longitude %% 360 # convert longitude
nino$hemisphere= nino$latitude < 0.0 # Hemisphere
nino$hemisphere= ifelse(nino$hemisphere, 'South', 'North')
nino$date=as.Date(as.character(nino$date),format="%y %m %d")#fix date format
data=nino[, !(names(nino) %in% c('humidity'))] # sub-data
# Remove rows with NA values in 'Air Temp' or 'Sea Surface Temp'
data=data[complete.cases(data$`air temp.`, data$s.s.temp.,data$zon.winds,data$mer.winds),]
summary(data) # check the data
# correlation plot
correlation=cor(subset(data,select=-c(date,hemisphere)),use="complete.obs")
corrplot(correlation, method="number")
# temperature vs position plots
a=ggplot(data,aes(x=`air temp.`,y=s.s.temp.,colour = hemisphere)) + geom_point(position = position_dodge(width = 0.4)) +xlab("Air.Temp") +
  ylab("sea surface temp") +
  ggtitle("plot 1 of Air.Temp vs sea surface temperature")
b=ggplot(data,aes(x=longitude,y=s.s.temp.)) + geom_point(position = position_dodge(width = 0.4)) +xlab("logitude") +
  ylab("sea surface temp") +
  ggtitle("plot 2 of longitude vs sea surface temperature")
c=ggplot(data,aes(x=longitude,y=`air temp.`)) + geom_point(position = position_dodge(width = 0.4)) +xlab("logitude") +
  ylab("air temp") +
  ggtitle("plot 3 of longitude vs air temperature")
grid.arrange(a,b,c, ncol=2)
nino94=nino[nino$year==94,]
nino94$time=nino94$month + nino94$day/30
#time series plot
par(mfcol=c(2,2),mar=c(4,4,2,1))
plot(nino94$time, nino94$`air temp.` ,type="l",lty=2, xlab="Time",ylab="Air temp",col="red")
plot(nino94$time, nino94$s.s.temp. ,type="l",lty=2, xlab="Time",ylab="Sea surface temp",col="orange")
plot(nino94$time, nino94$humidity ,type="l",lty=2, xlab="Time",ylab="Humidity",col="yellow")
plot(nino94$time, nino94$zon.winds ,type="l",lty=2, xlab="Time",ylab="Zonal wind",col="blue")
```