Problem Set 4

Use the 2014 daily GHCN temperature data to obtain monthly means of the daily maximum temperature, and of the daily minimum temperature. Compute the ranges of these two sets of 12 values for each station.

SAS Code as follows:

##Set filename for file downloaded

libname mydata '~/506/';

filename ghcnd\_gz pipe "gzip -dc 2014.csv.gz" lrecl=80;

##Read in temperature data and format for use

data ghcnd;

infile ghcnd\_gz delimiter=",";

input station $ date : yymmdd8. obstype $ obsval;

format date mmddyy10.;

month = month(date);

if obstype="TMAX" or obstype="TMIN";

obsval=obsval/10;

##Read in station data for use, exclude unused colomns

data stations;

infile "ghcnd-stations.txt";

input station $ 1-11 lat 13-20 lon 22-30 elev 32-37 state $ 39-40;

proc sort data=ghcnd out=ghcnd2;

by station;

proc sort data=stations out=stations2;

by station;

data ghcnd3;

merge ghcnd2(in=x) stations2(in=y);

by station;

if x=1 and y=1;

##Compute mean temperature for each station for each month

proc summary data=ghcnd3 nway;

class station month obstype;

output out=ghcnd4

mean(obsval)=mntmp;

##Use the temperature type as variables

proc transpose data=ghcnd4(drop=\_TYPE\_ \_FREQ\_) out=ghcnd5;

by station month;

id obstype;

var mntmp;

##Compute the range

data ghcnd6(drop=\_NAME\_ elev state);

merge ghcnd5(in=x) stations2(in=y);

by station;

if x=1 and y=1;

tmprange=TMAX-TMIN;

##Print out the data for review

proc print data=ghcnd6;

##Save the final data

data mydata.finaldata;

set ghcnd6;

run;

1. Calculate the difference of the two ranges for each station (i.e. the range of maximum temperatures minus the range of minimum temperatures). Identify the stations which have the least and greatest values for this difference.

##Using the data we get from preparation.

libname mydata '~/506/';

data df;

set mydata.finaldata;

##Use sql to get the range

proc sql;

create table tmp\_range as

select station,

range(TMAX) as tmax\_range,

range(TMIN) as tmin\_range,

range(TMAX)-range(TMIN) as range\_diff

from df

group by station;

quit;

proc print data=tmp\_range;

##Get the stations with the max and min range.

proc summary data=tmp\_range;

output out=maxrange(drop=\_TYPE\_ \_FREQ\_)

maxid(range\_diff(station))=station

max(range\_diff)=max\_range;

proc summary data=tmp\_range;

output out=minrange(drop=\_TYPE\_ \_FREQ\_)

minid(range\_diff(station))=station

min(range\_diff)=min\_range;

proc print data=maxrange;

proc print data=minrange;

data mydata.tmp\_range;

set tmp\_range;

run;

We got the following stations with the max and min temperate range.

Obs station max\_range

1 CA006059 32.0495

Obs station min\_range

1 USS0006H -266.395

1. In SAS, produce a reduced dataset with indicator variables indicating the stations that are in the bottom 10% of the distribution of values for each of the two ranges. Drop the stations that are not in the bottom 10% for either range. Export the indicator variables and the geographic coordinates of these stations to a text file using proc export.

libname mydata '~/506/';

data tmp\_range;

set mydata.tmp\_range;

## Get the quantiles for 10%

proc univariate data=tmp\_range;

var tmax\_range tmin\_range;

output out=range10 pctlpts=10 pctlpre=tmax\_range tmin\_range;

proc print range10;

data stations;

infile "ghcnd-stations.txt";

input station $ 1-11 lat 13-20 lon 22-30 elev 32-37 state $ 39-40;

proc sort data=stations out=stations2;

by station;

data df1;

merge tmp\_range(in=x) stations2(in=y);

by station;

if x=1 and y=1;

## Indicating for each row. Delete the ones in neither bottom 10%.

data df2;

set df1;

if tmax\_range ge 6.68495 and tmin\_range ge 6.54762 then delete;

else if tmax\_range ge 6.68495 and tmin\_range lt 6.54762 then indicator=2;

else if tmax\_range lt 6.68495 and tmin\_range ge 6.54762 then indicator=3;

else indicator=1;

##Export the data to a txt file for later analysis in R

proc export data=df2 dbms=tab outfile='bottom10.txt' replace;

run;

Load these coordinates into R, and produce a map showing where these stations are located on a map of the Earth. Use three different colors so it is clear which stations are in the bottom 10% of the distribution for either range value, or for both of them. Briefly comment on your findings.

df=read.table(file="C:/Users/Heathtasia/Desktop/506/bottom10.txt",header=TRUE,sep="\t")

df$indicator=as.factor(df$indicator)

levels(df$indicator)=c("both bottom 10%","tmin range bottom 10%","tmax range bottom 10%")

library(mapproj)

library(maptools)

coord=mapproject(df$lon,df$lat)

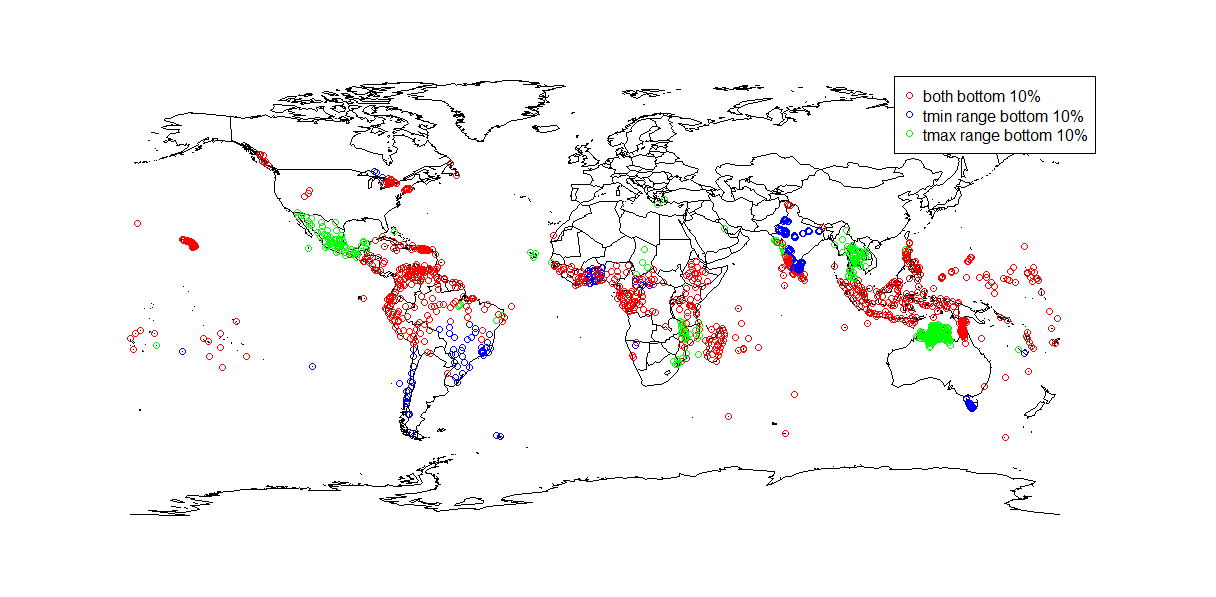
map()

##Draw points on the map

points(coord,col=c("red","blue","green")[df$indicator])

legend(x="topright", legend = levels(df$indicator), col=c("red","blue","green"), pch=1)

We get the map of temperature range in the bottom 10% as follows:



From the map we find that the regions near the equator has both tmax range and tmin range in the bottom 10%, which means the temperature is more stable in those areas.

And also a majority of areas near the ocean have either more stable tmin or tmax. The ocean play an important part in the climate nearby.

2. Using the 2014 daily maximum GHCN temperature data, calculate the mean and standard deviation of the values within each month for each station.

libname mydata '~/506/';

filename ghcnd\_gz pipe "gzip -dc 2014.csv.gz" lrecl=80;

data ghcnd(rename=(obsval=tmax));

infile ghcnd\_gz delimiter=",";

input station $ date : yymmdd8. obstype $ obsval;

format date mmddyy10.;

month = month(date);

if obstype="TMAX";

obsval=obsval/10;

##Get the mean and std of tmax

proc sql;

creat table tmax\_dist as

select station,month, mean(tmax) as mntmax, std(tmax) as stdtmax

from ghcnd

group by station,month;

quit;

##Export the data to a txt file for later analysis in R

proc export data = tmax\_dist outfile = "tmax\_dist.txt" dbms=tab replace;

run;

##Read in data in R

df2=read.table(file="C:/Users/Heathtasia/Desktop/506/tmax\_dist.txt",header=TRUE,sep="\t")

##Delete the missing values

df3=na.omit(df2)

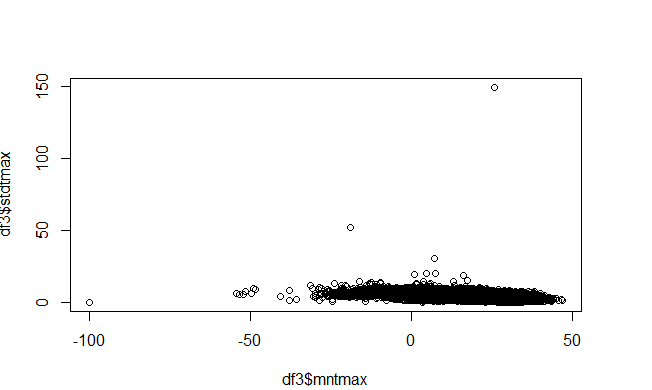
##Calculate the simple correlation of mean and std of tmax

cor(df3$mntmax,df3$stdtmax)

#-0.5063753 the mean and std of tmax are negatively correlated

## Look at a simple scatter plot

plot(df3$mntmax,df3$stdtmax)



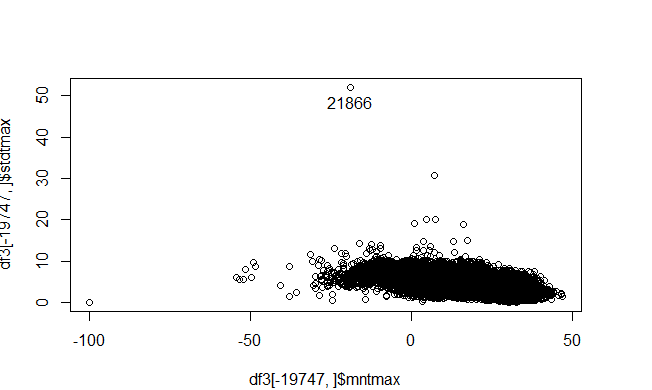
## Identify outliers

identify(df3$mntmax,df3$stdtmax)

#[1] 19747

## Remove the outlier to see more in detail

plot(df3[-19747,]$mntmax,df3[-19747,]$stdtmax)



3. Use the GHCN data to obtain the 2013 and 2014 mean values for daily maximum temperature for each day in January at each station. Calculate the difference between these two means (e.g. 2014 January mean minus 2013 January mean).

libname mydata '~/506/';

filename ghcnd\_gz pipe "gzip -dc 2014.csv.gz" lrecl=80;

data ghcnd14(rename=(obsval=tmax14));

infile ghcnd\_gz delimiter=",";

input station $ date : yymmdd8. obstype $ obsval;

format date mmddyy10.;

month = month(date);

if obstype="TMAX";

obsval=obsval/10;

if month=1;

day=day(date);

keep station day obsval;

##Calculate mean of tmax for each day in Jan.

proc sql;

create table mntmax\_14 as

select station,day,mean(tmax14) as mntmax14

from ghcnd14

group by station,day;

quit;

filename ghcnd\_gz pipe "gzip -dc 2013.csv.gz" lrecl=80;

data ghcnd13(rename=(obsval=tmax13));

infile ghcnd\_gz delimiter=",";

input station $ date : yymmdd8. obstype $ obsval;

format date mmddyy10.;

month = month(date);

if obstype="TMAX";

obsval=obsval/10;

if month=1;

day=day(date);

keep station day obsval;

proc sql;

create table mntmax\_13 as

select station,day,mean(tmax13) as mntmax13

from ghcnd13

group by station,day;

quit;

data meantmax;

merge mntmax\_14(in=x) mntmax\_13(in=y);

by station day;

if x=1 and y=1;

tmax\_diff=mntmax14-mntmax13;

##Export the data to a txt file for later analysis in R

proc export data=meantmax dbms=tab outfile="meantmax.txt" replace;

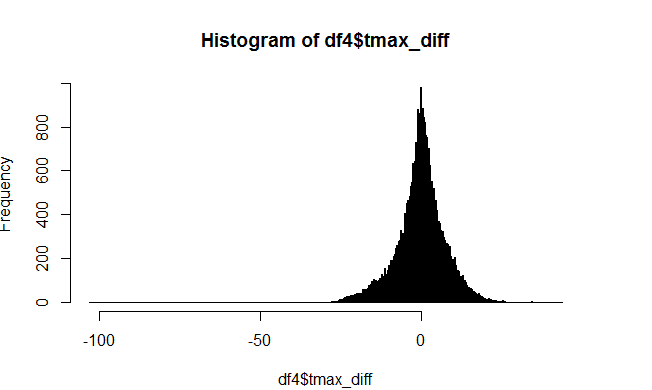
run;

Then (i) export all the mean differences to R and make a histogram of them, (ii) export the geographic coordinates of the stations in the top and bottom 10% of the distribution, then use R to make a map of these two sets of points. Briefly comment on your findings.

(i)

df4=read.table(file="C:/Users/Heathtasia/Desktop/506/meantmax.txt",header=TRUE,sep="\t")

hist(df4$tmax\_diff,breaks=1000)



(ii)

##Calculate the quantiles

proc univariate data=meantmax;

var tmax\_diff;

output out=tmax\_diff\_quantile pctlpts=10,90 pctlpre=tmax\_diff;

proc print data=tmax\_diff\_quantile;

##Get the bottom 10% and top 10% with indicators

data meantmax\_10;

set meantmax;

if tmax\_diff ge -9.366667 and tmax\_diff lt 8.904706 then delete;

else if tmax\_diff lt -9.366667 then indicator=1;

else if tmax\_diff ge 8.904706 then indicator=2;

data stations;

infile "ghcnd-stations.txt";

input station $ 1-11 lat 13-20 lon 22-30 elev 32-37 state $ 39-40;

proc sort data=stations out=stations2;

by station;

data meantmax\_10\_map;

merge meantmax\_10(in=x) stations2(in=y);

by station;

if x=1 and y=1;

##Export the data to a txt file for later analysis in R

proc export data=meantmax\_10\_map dbms=tab outfile="meantmaxmap.txt" replace;

In R we plot the map for the top 10% and bottom 10% for tmax\_diff

df5=read.table(file="C:/Users/Heathtasia/Desktop/506/meantmaxmap.txt",header=TRUE,sep="\t")

df5$indicator=as.factor(df5$indicator)

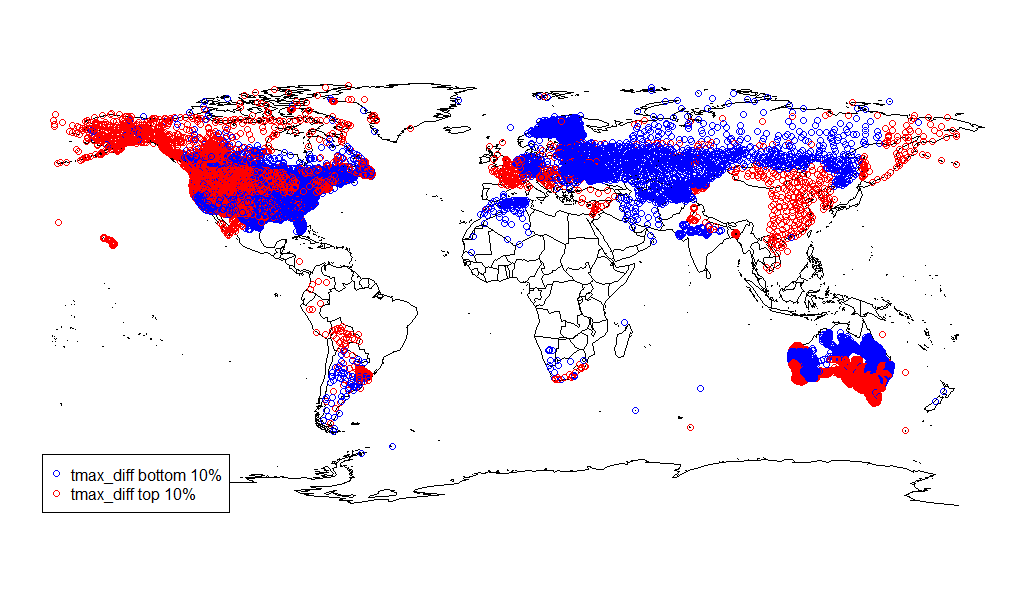
levels(df5$indicator)=c("tmax\_diff bottom 10%","tmax\_diff top 10%")

map()

coord=mapproject(df5$lon,df5$lat)

points(coord,col=c("blue","red")[df5$indicator])

legend(x="bottomleft", legend = levels(df5$indicator), col=c("blue","red"), pch=1)



From the map we know that the areas with change(from 2013 to 2014) in tmax in the top 10% and bottom 10% are away from the equator and the polar. The Africa seems to have less extreme tmax changes from 2013 to 2014. The blue ones are the bottom 10% tmax changing area and the red ones are the top 10% tmax changing area.