

# STATSc173 PROJECT

Xuanyao Qian SID: 905-723-084

2/29/2024

## ***Introduction***

### **-General Background-**

The data contains weather data from 555 unique weather stations in India. For the purpose of this project, I have selected three variables: cloud coverage as the response variable, humidity and temperature as explanatory variable. Each variable was grouped by the weather station and average was taken.

The dataset was originated from Kaggle. The data is accessible through the link:<https://www.kaggle.com/datasets/nelgiriyewithana/indian-weather-repository-daily-snapshot>

## ***Data Description***

### **-Loading Data-**

```
library(prettydoc)
```

```
## Warning: package 'prettydoc' was built under R version 4.3.2
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
library(geoR)
```

```
## Warning: package 'geoR' was built under R version 4.3.2
```

```
## -----
## Analysis of Geostatistical Data
## For an Introduction to geoR go to http://www.leg.ufpr.br/geoR
## geoR version 1.9-3 (built on 2023-12-11) is now loaded
## -----
```

```
library(gstat)
```

```
## Warning: package 'gstat' was built under R version 4.3.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':  
##  
##     filter, lag
```

```
## The following objects are masked from 'package:base':  
##  
##     intersect, setdiff, setequal, union
```

```
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.3.2
```

```
library(sp)
```

```
## Warning: package 'sp' was built under R version 4.3.2
```

```
data<-read.csv('C:/Users/PaFuu/Desktop/IndianWeatherRepository.csv')  
  
data_n <- data %>%  
  group_by(longitude,latitude)%>% #group by 555 unique weather stations  
  summarise('avg_pressure' = median(pressure_mb), 'avg_humidity' = median(humidity), 'avg_cloud_coverage' = median(cloud_coverage))
```

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

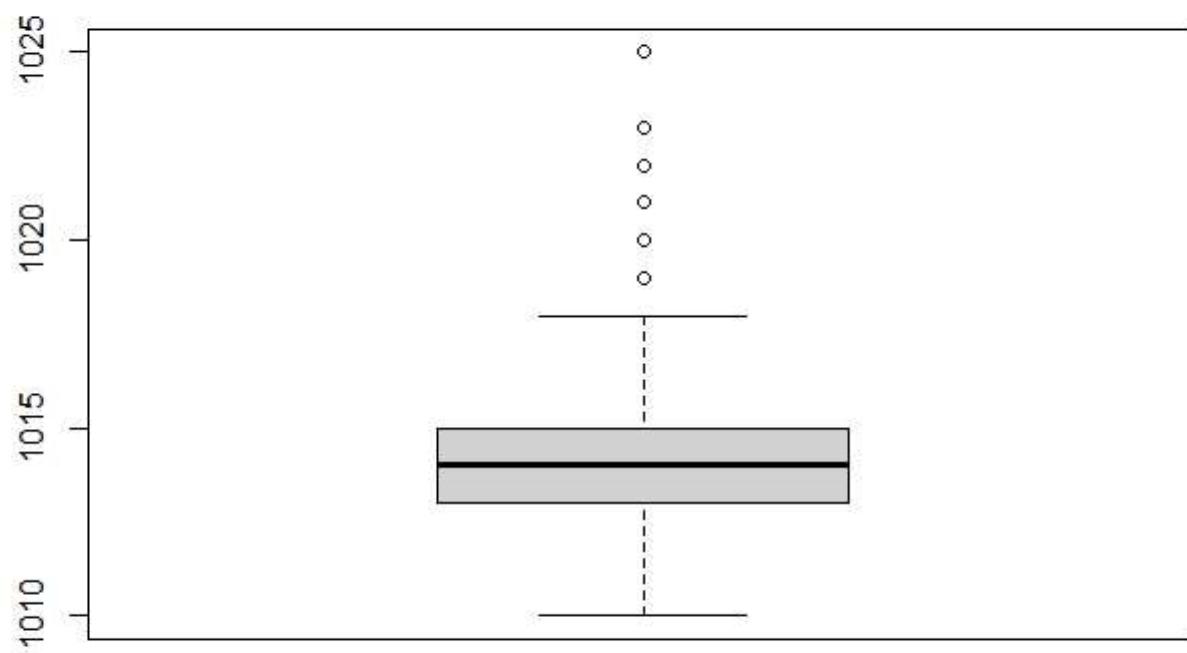
```
#variable chosen: average temperature, humidity, cloud coverage in each station  
print(data_n)
```

```
## # A tibble: 555 × 5  
## # Groups:   longitude [435]  
##   longitude latitude avg_pressure avg_humidity avg_cloud_coverage  
##       <dbl>     <dbl>      <dbl>      <dbl>            <dbl>  
## 1     69.0     22.2      1013       68             3  
## 2     69.6     21.6      1013       60             3  
## 3     70.1     22.5      1013       54             3  
## 4     70.4     20.9      1012       64             3  
## 5     70.5     21.5      1013       54             3  
## 6     70.8     22.3      1013       45             3  
## 7     70.8     22.8      1013       47             3  
## 8     70.9     26.9      1014       29             3  
## 9     71.0     20.7      1012       61             3  
## 10    71.0     20.8      1012       58             3  
## # i 545 more rows
```

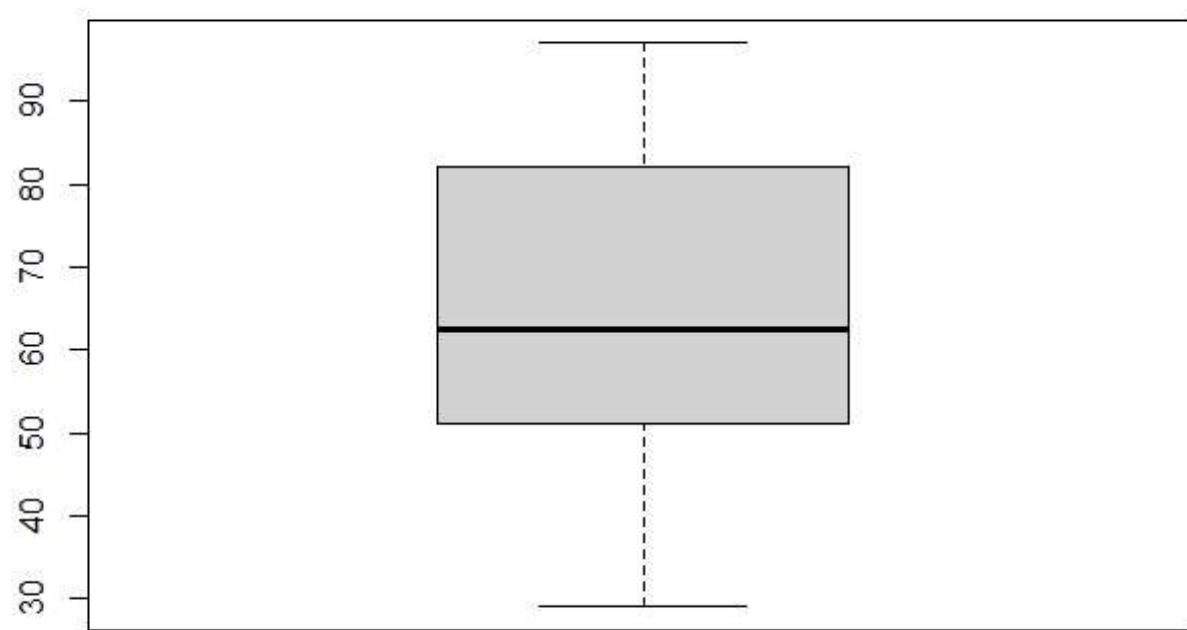
## Analyzing the raw data

Both boxplot and histogram show that the target variable `cloud_coverage` is highly skewed with many outliers.

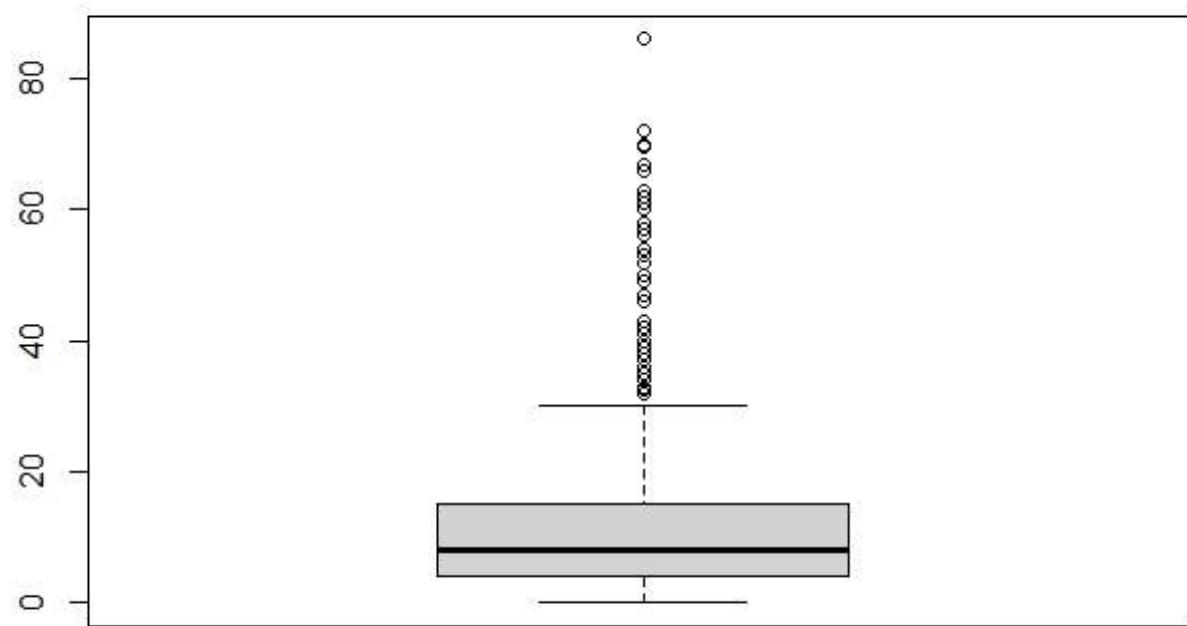
```
boxplot(data_n$avg_pressure)
```



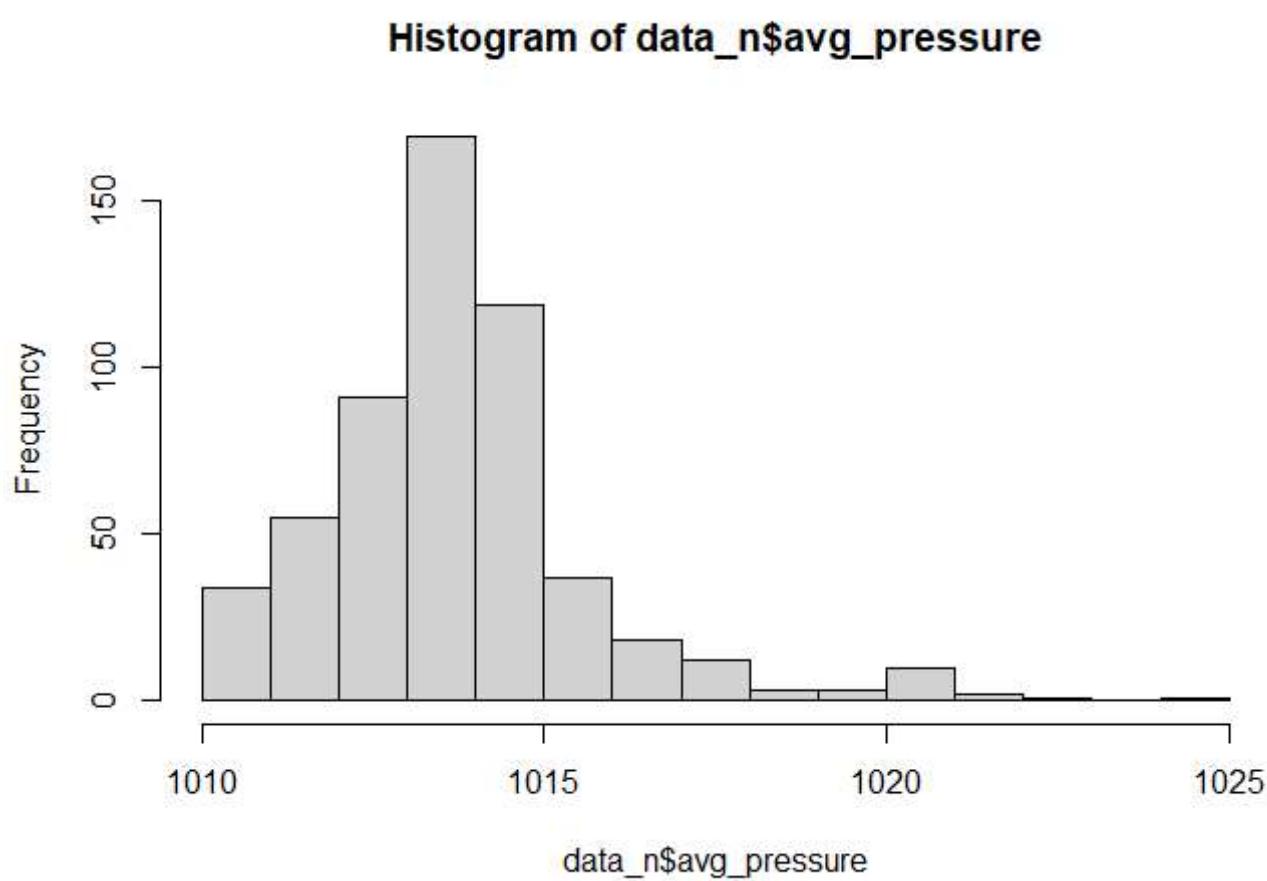
```
boxplot(data_n$avg_humidity)
```



```
boxplot(data_n$avg_cloud_coverage)
```

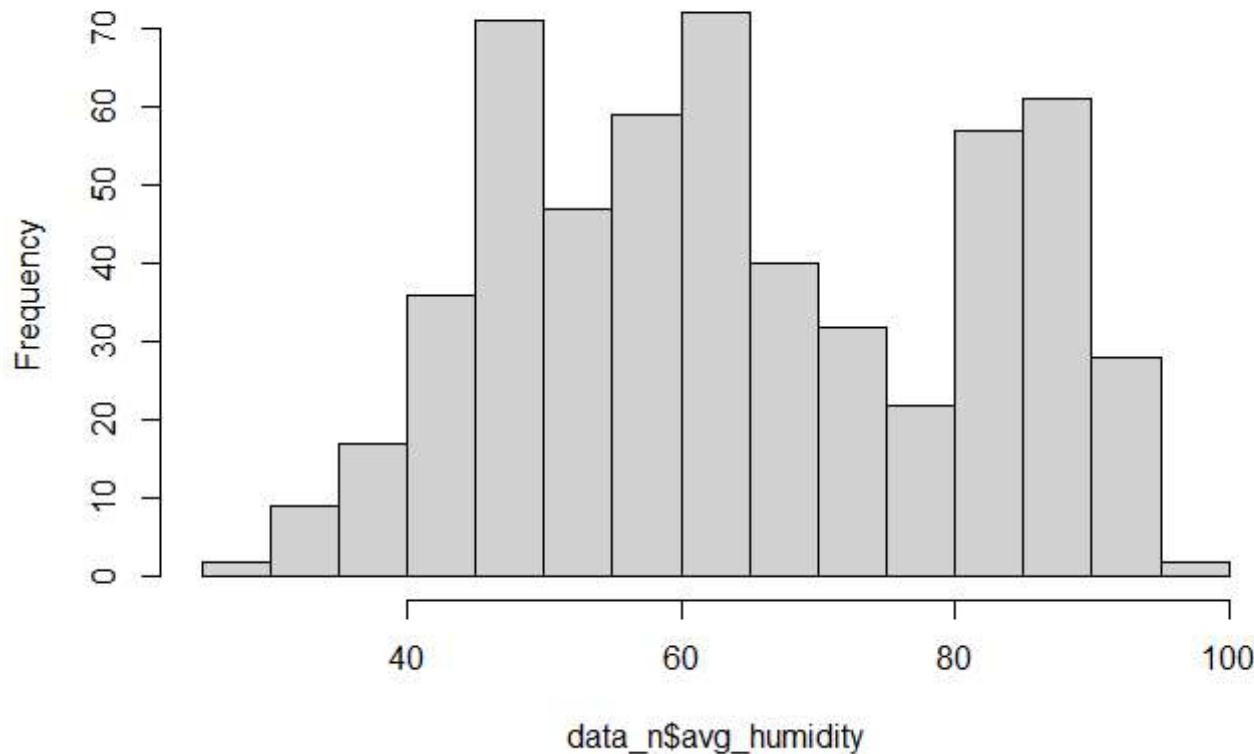


```
hist(data_n$avg_pressure)
```



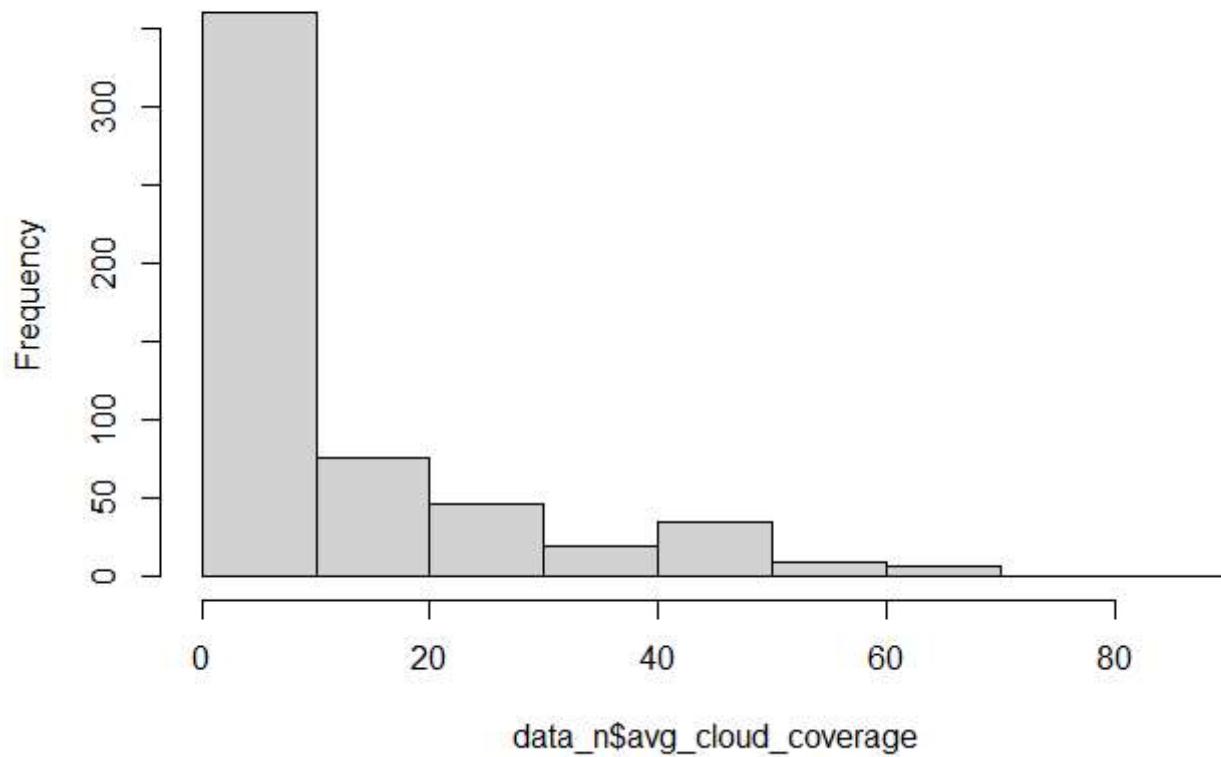
```
hist(data_n$avg_humidity)
```

**Histogram of data\_n\$avg\_humidity**



```
hist(data_n$avg_cloud_coverage)
```

**Histogram of data\_n\$avg\_cloud\_coverage**

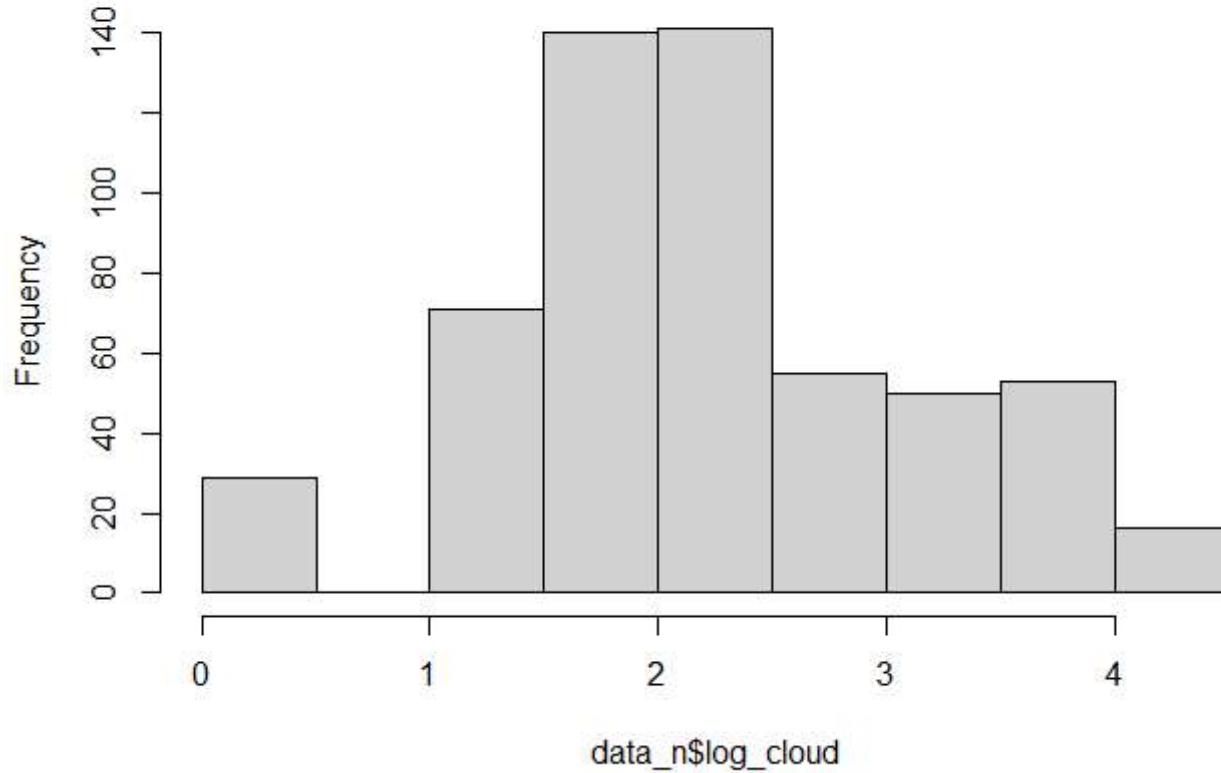


#### -Log Transformation-

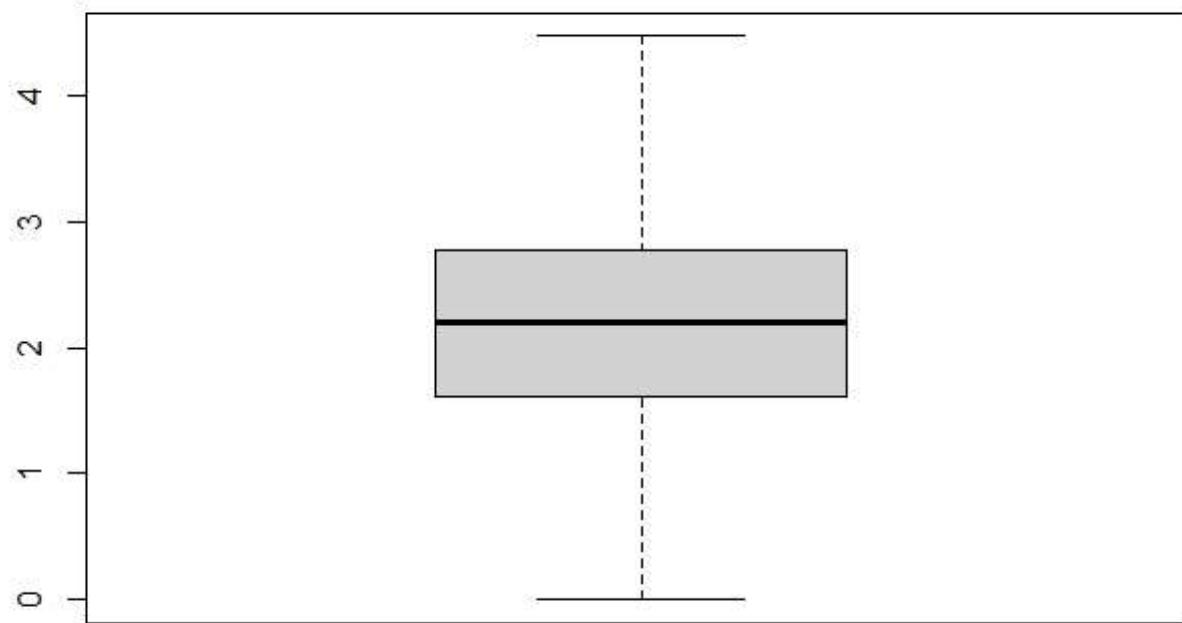
In order to make the distribution of target variable more normal, we will perform a log transformation of `avg_cloud`. After the transformation, we could see that the normality of our target variable is greatly improved.

```
data_n$avg_cloud_coverage<-data_n$avg_cloud_coverage+1  
data_n<-data_n %>%  
  mutate("log_cloud" = log(avg_cloud_coverage))  
hist(data_n$log_cloud)
```

Histogram of data\_n\$log\_cloud



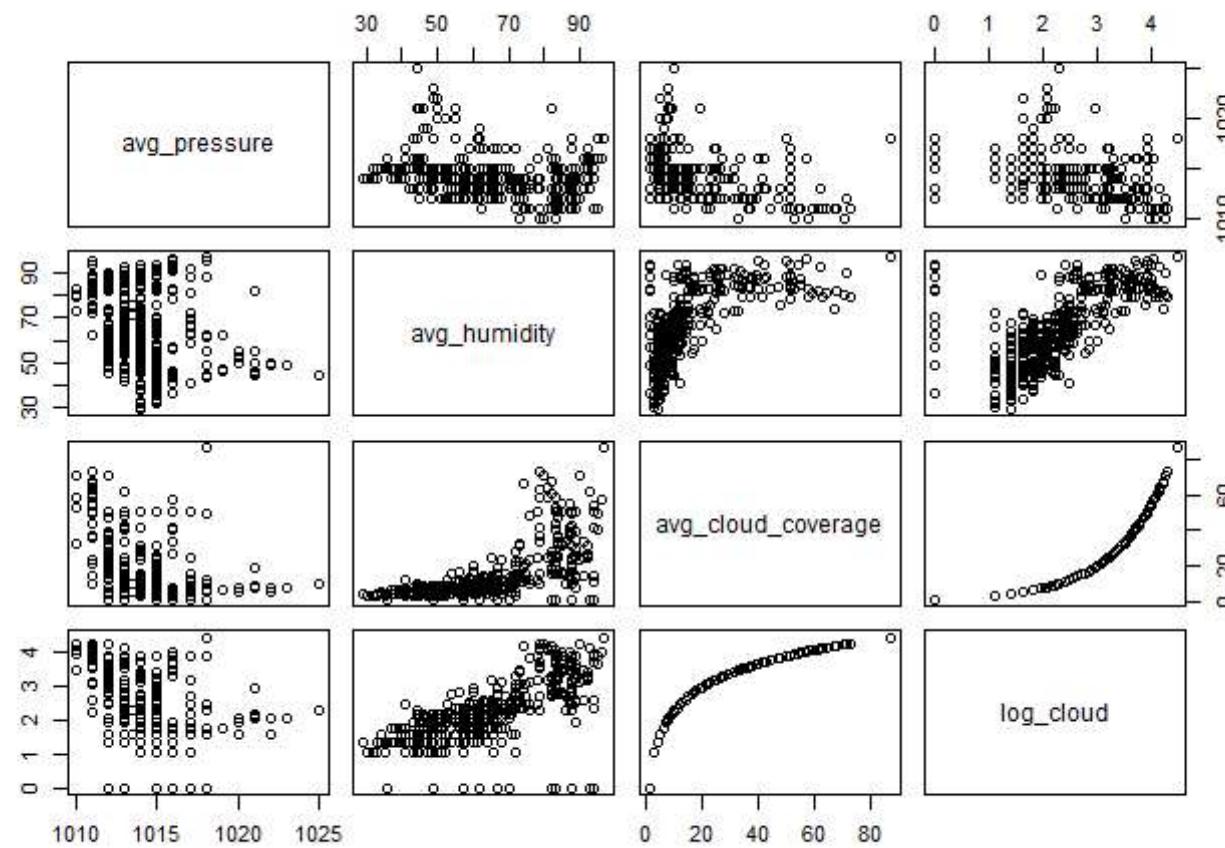
```
boxplot(data_n$log_cloud)
```



### -Correlation-

Inspecting the correlation, there is a `-0.37` correlation between `log_cloud` and `avg_pressure`, and a `0.56` correlation between `log_cloud` and `avg_humidity`, indicating that we might use co-kriging for better prediction.

```
pairs(data_n[,3:6])
```



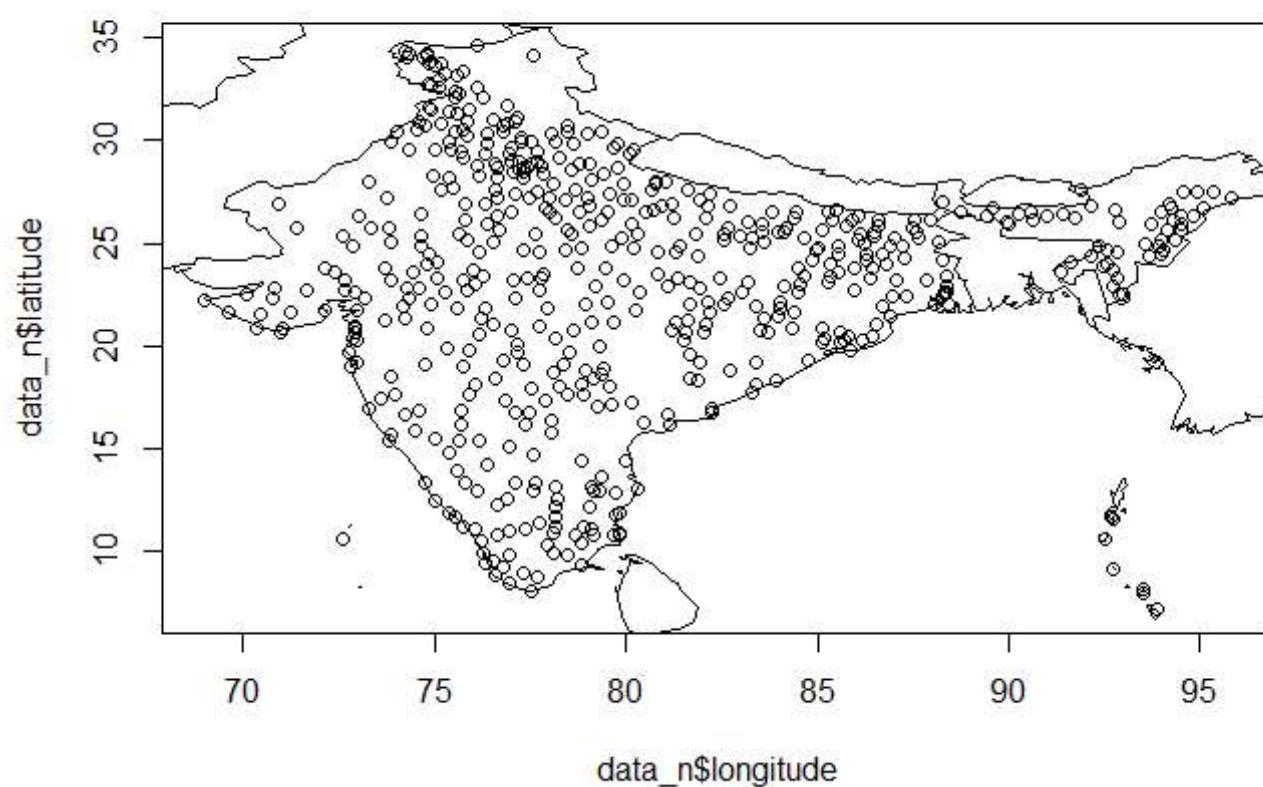
```
cor(data_n[,3:6])
```

```
##                                     avg_pressure avg_humidity avg_cloud_coverage log_cloud
## avg_pressure                 1.0000000   -0.2550064      -0.3708578 -0.3743522
## avg_humidity                -0.2550064    1.0000000       0.6412432  0.5618926
## avg_cloud_coverage          -0.3708578     0.6412432      1.0000000  0.8718418
## log_cloud                   -0.3743522     0.5618926       0.8718418  1.0000000
```

## Visualizing the raw data

-Plot Data Points on the World Map-

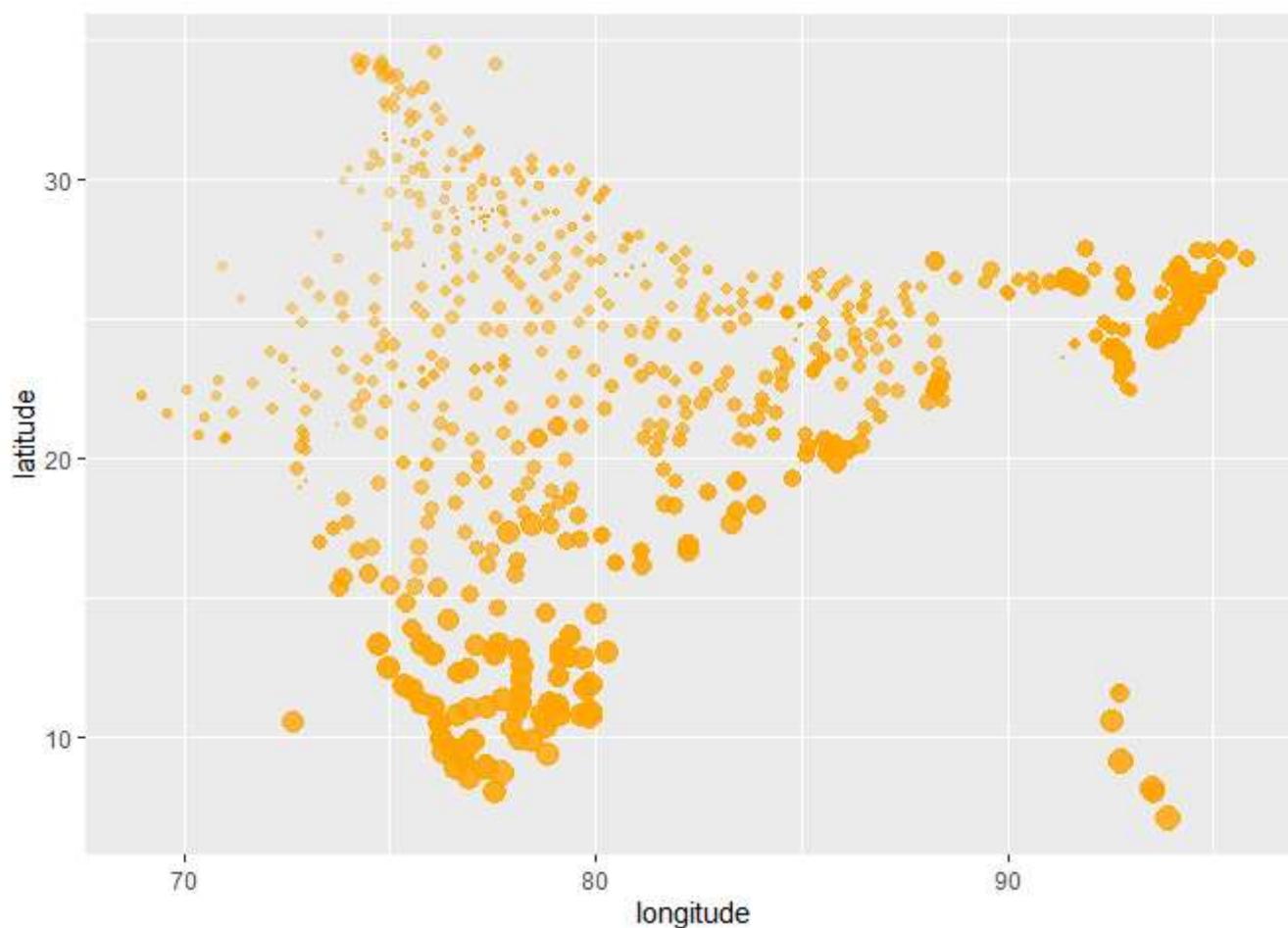
```
plot(data_n$longitude,data_n$latitude)
map("world",add = T)
```



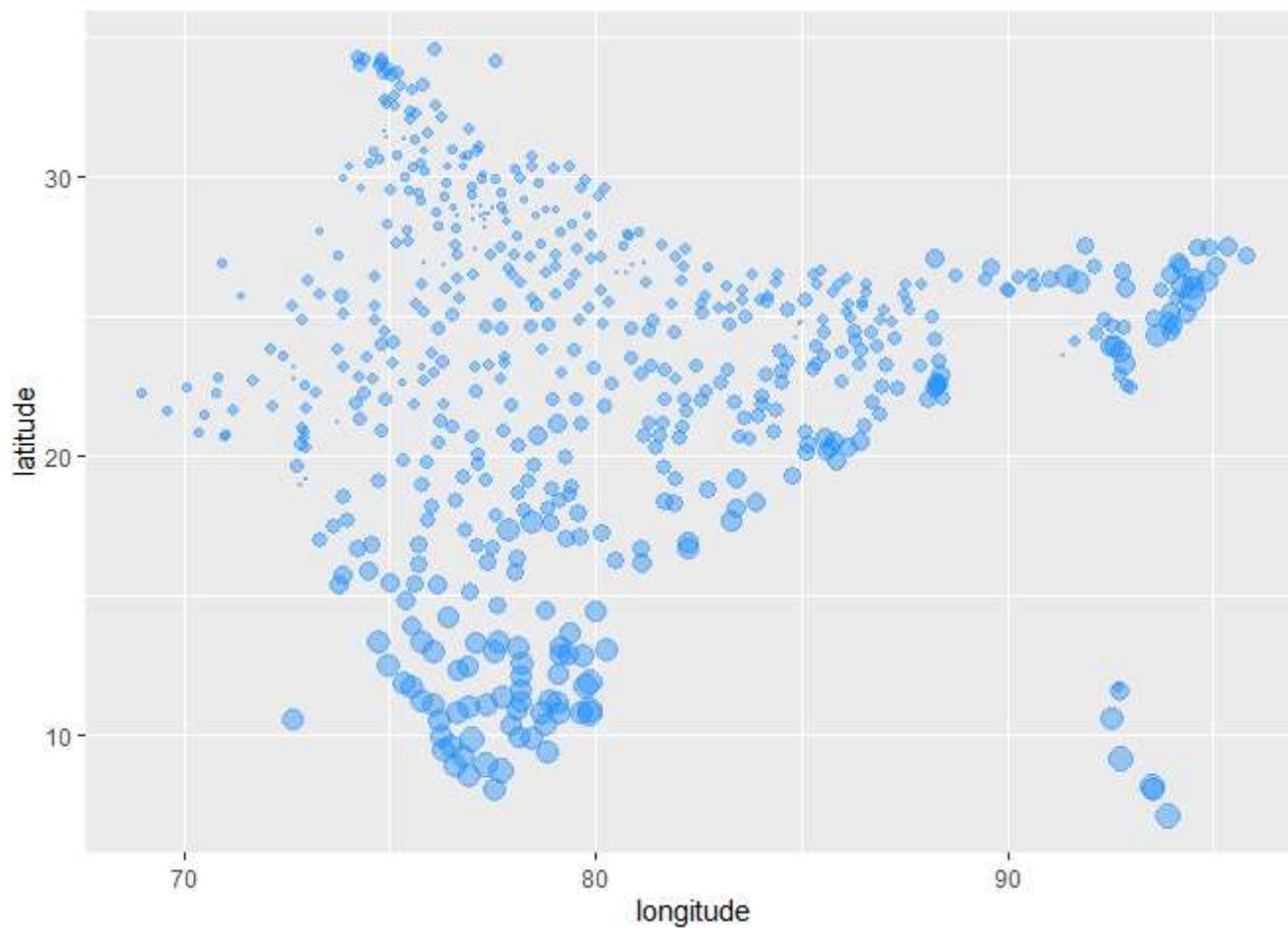
-Visualizing Correlation-

Visualizing correlation on the map, we could see a clear trend that higher temperature tend to have bigger cloud coverage, while the trend with pressure is not so clear.

```
#plot cloud coverage against temperature  
ggplot(data_n,aes(x=longitude,y=latitude))+geom_point(size=data_n$log_cloud,alpha=data_n$avg_humidity)
```



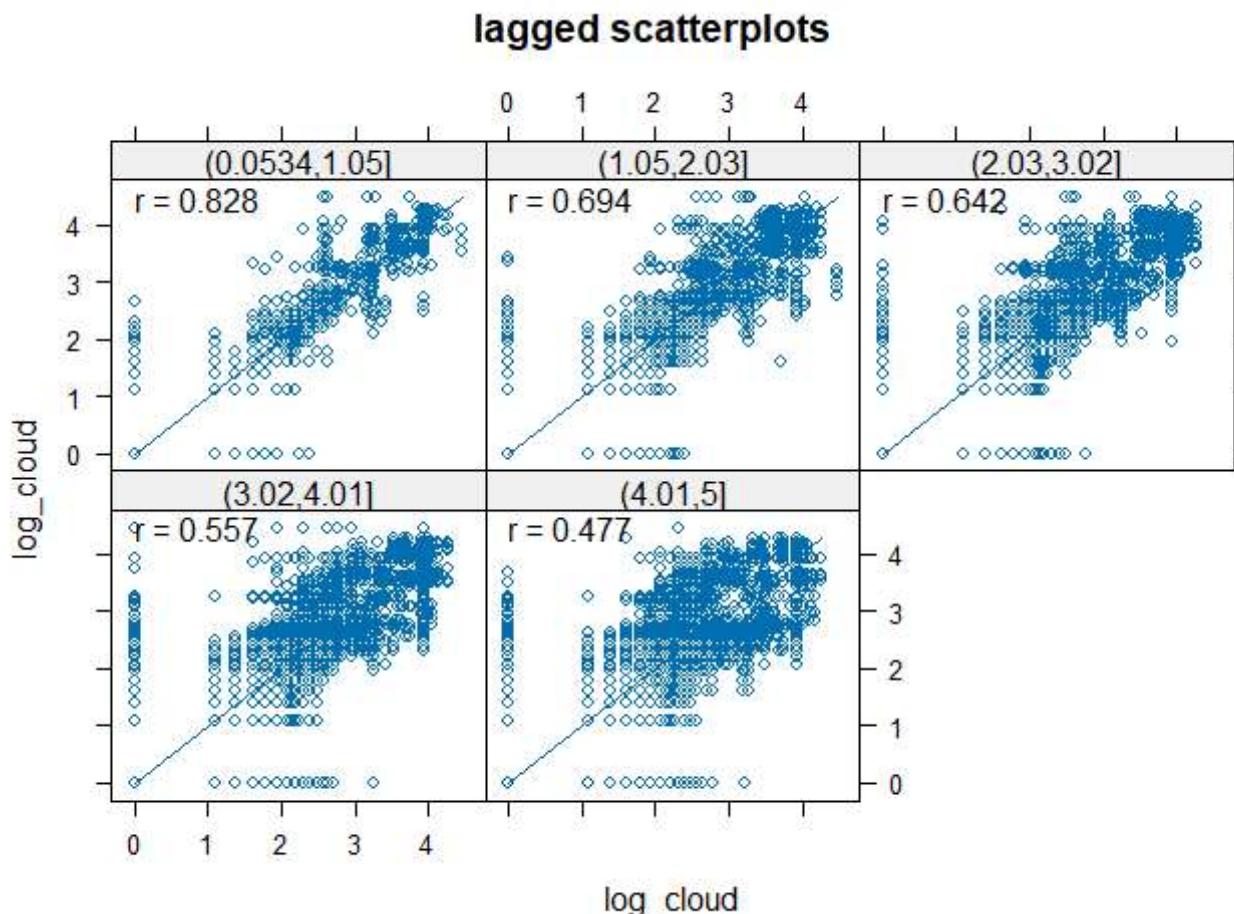
```
#plot cloud coverage against pressure  
ggplot(data_n,aes(x=longitude,y=latitude))+geom_point(size=data_n$log_cloud,alpha=data_n$avg_pressure)
```



### -h-scatterplot-

The h-scatter plot become more sparsely distributed as distance increases, indicating spatial correlation exists.

```
data_target <- data_n[,c(1,2,6)]  
#creating the h-scatter plot  
coordinates(data_target)<-~longitude+latitude  
qq<- hscat(log_cloud~1,data_target,5)  
plot(qq)
```



## Variograms

### -Variograms Using geoR-

Next, we conduct initial analysis by creating a semi-variogram. By fitting three different models, we could see that fit1 with Gaussian distribution fits the best.

```
#creating sample variogram using geoR and fit different models
geo_data <- as.geodata(data_target)
variogram_1 <- variog(geo_data,max.dist = 20)
```

```
## variog: computing omnidirectional variogram
```

```
plot(variogram_1)
ini.values <- expand.grid(seq(0,3,by=.05),seq(20,30,by=1))
fit1 <- variofit(variogram_1,cov.model = 'gau',ini.cov.pars = ini.values,fix.nugget = T,nugget=0.2)

## variofit: covariance model used is gaussian
## variofit: weights used: npairs
## variofit: minimisation function used: optim
## variofit: searching for best initial value ... selected values:
##           sigmasq phi   tausq kappa
## initial.value "3"    "21"   "0.2" "0.5"
## status        "est"   "est"  "fix"  "fix"
## loss value: 642.11280703351
```

```
fit2 <- variofit(variogram_1,cov.model = 'sph',ini.cov.pars = ini.values,fix.nugget = T,nugget=0.2)
```

```
## variofit: covariance model used is spherical
## variofit: weights used: npairs
## variofit: minimisation function used: optim
## variofit: searching for best initial value ... selected values:
##           sigmasq phi   tausq kappa
## initial.value "1.6"   "30"   "0.2" "0.5"
## status        "est"   "est"  "fix"  "fix"
## loss value: 6920.53189563451
```

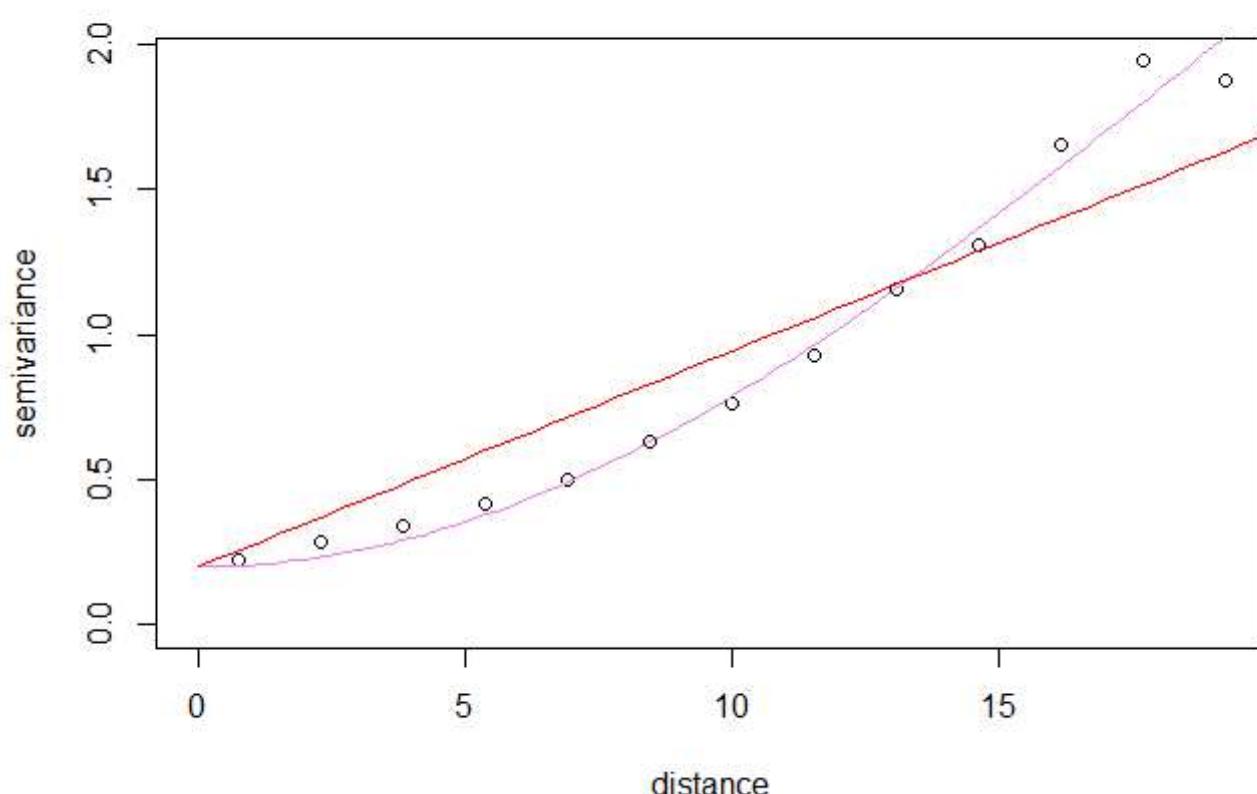
```
## Warning in variofit(variogram_1, cov.model = "sph", ini.cov.pars = ini.values,
## : unreasonable initial value for phi (too high)
```

```
fit3 <- variofit(variogram_1,cov.model = 'exp',ini.cov.pars = ini.values,fix.nugget = T,nugget=0.2)
```

```
## variofit: covariance model used is exponential
## variofit: weights used: npairs
## variofit: minimisation function used: optim
## variofit: searching for best initial value ... selected values:
##           sigmasq phi   tausq kappa
## initial.value "2.75" "30"  "0.2" "0.5"
## status        "est"  "est" "fix" "fix"
## loss value: 8015.61185564585
```

```
## Warning in variofit(variogram_1, cov.model = "exp", ini.cov.pars = ini.values,
## : unreasonable initial value for phi (too high)
```

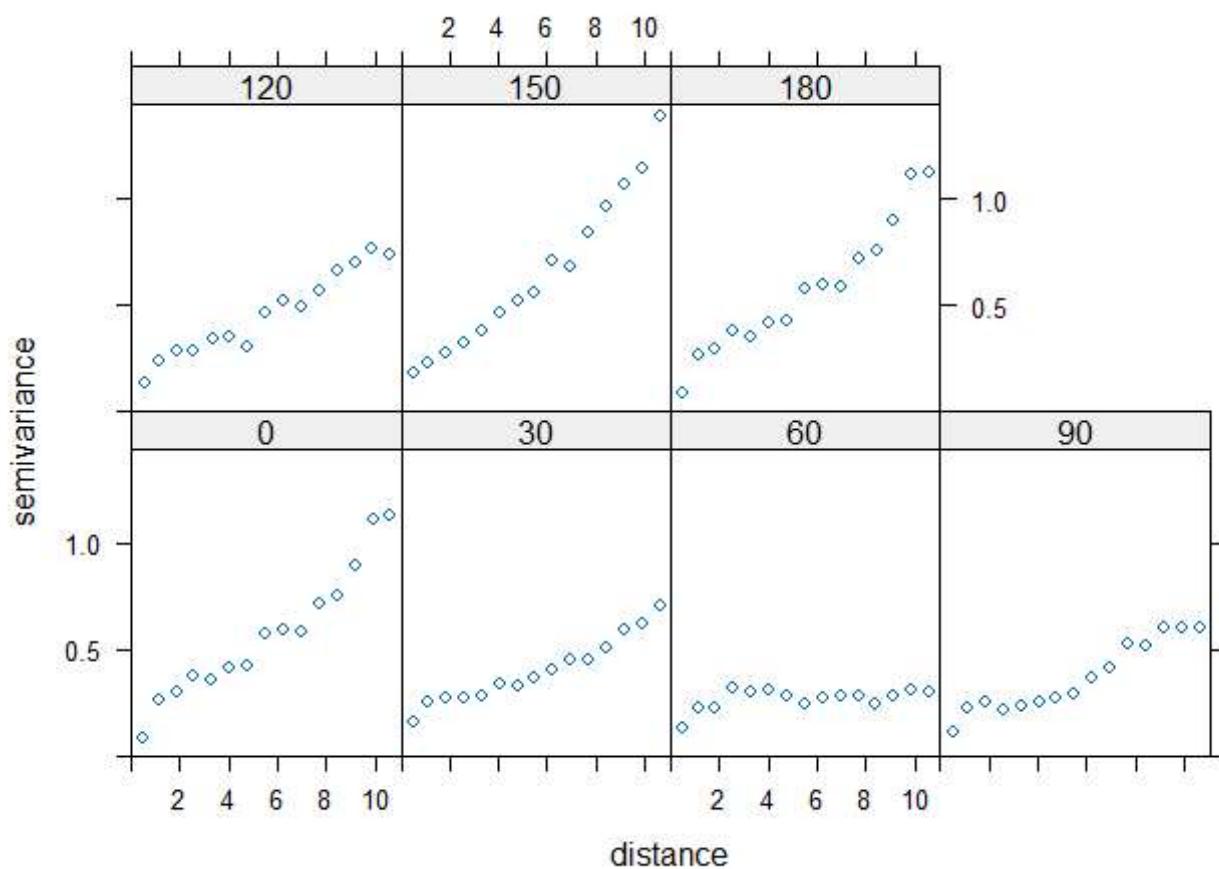
```
lines(fit1,col='violet')
lines(fit2,col='purple')
lines(fit3,col='red')
```



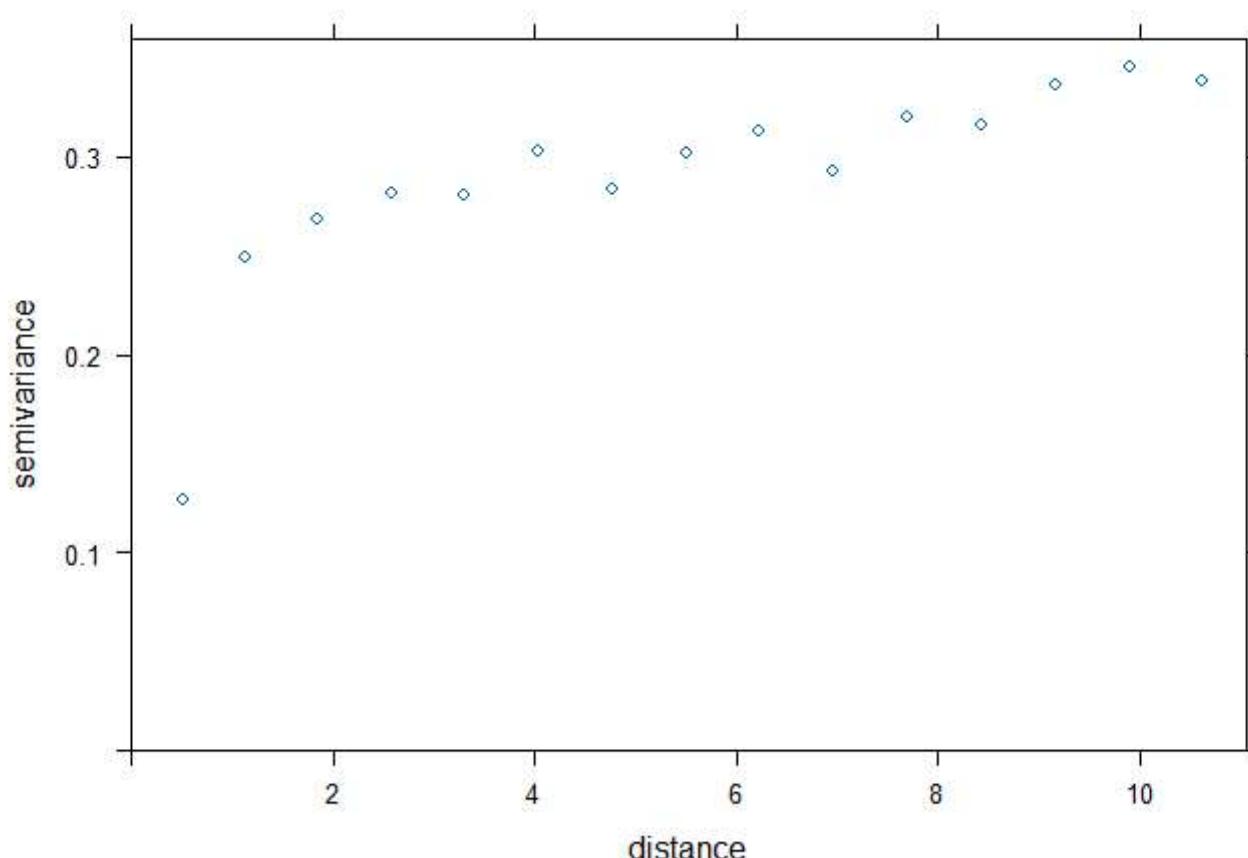
### -Variograms Using gstat-

Next, we create and fit the variogram again using gstat. As spatial correlation is suspected, we would include a trend.

```
data_target_gstat <- data_n[,c(1,2,6)]  
  
#create variogram based on different directions  
g <- gstat(id='log_cloud',formula = log_cloud~1,locations = ~longitude+latitude,data=data_target_gst  
vario_g<-variogram(g,cutoff=11,alpha=seq(0,180,30))  
plot(vario_g)
```

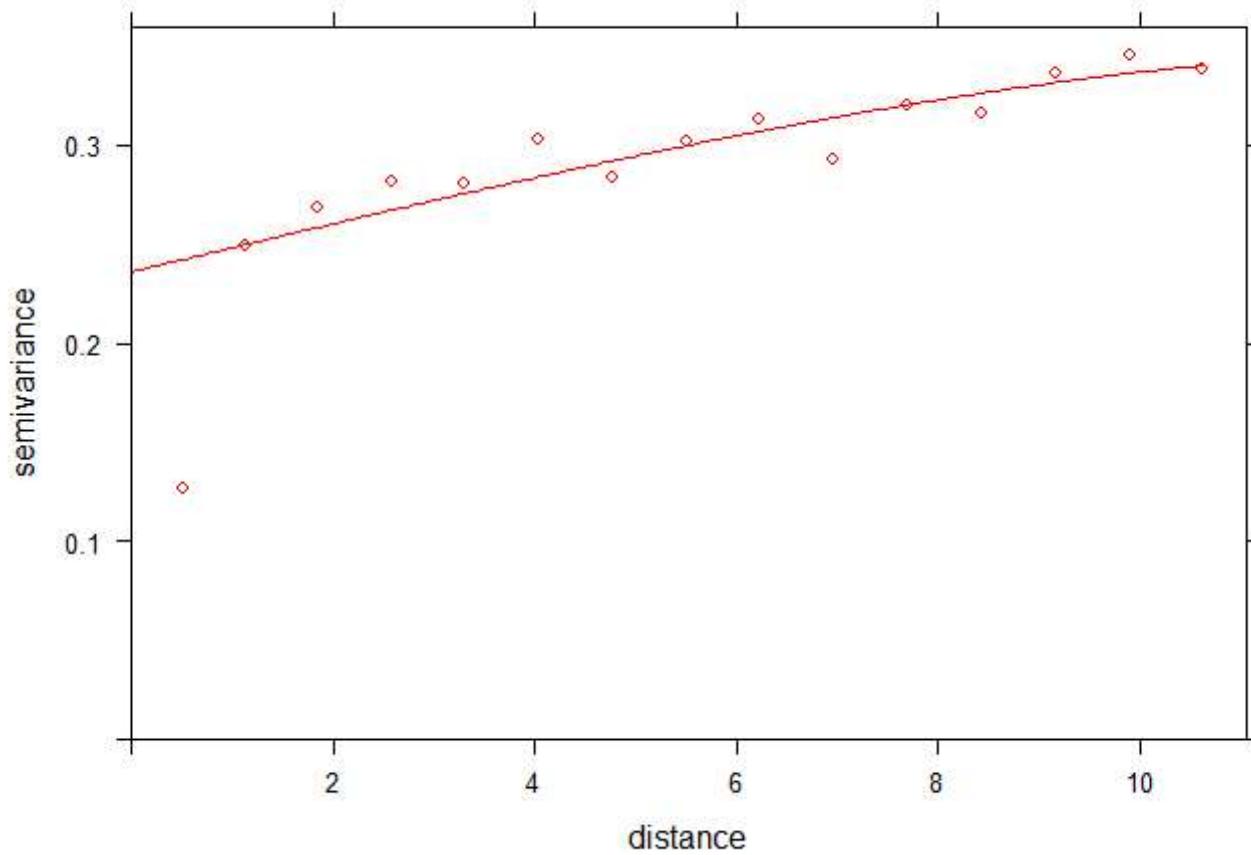


```
#create variogram with a trend
g_trend <- gstat(id='log_cloud',formula = log_cloud~longitude+latitude,locations = ~longitude+latitude)
vario_trend <-variogram(g_trend,cutoff=11)
plot(vario_trend)
```

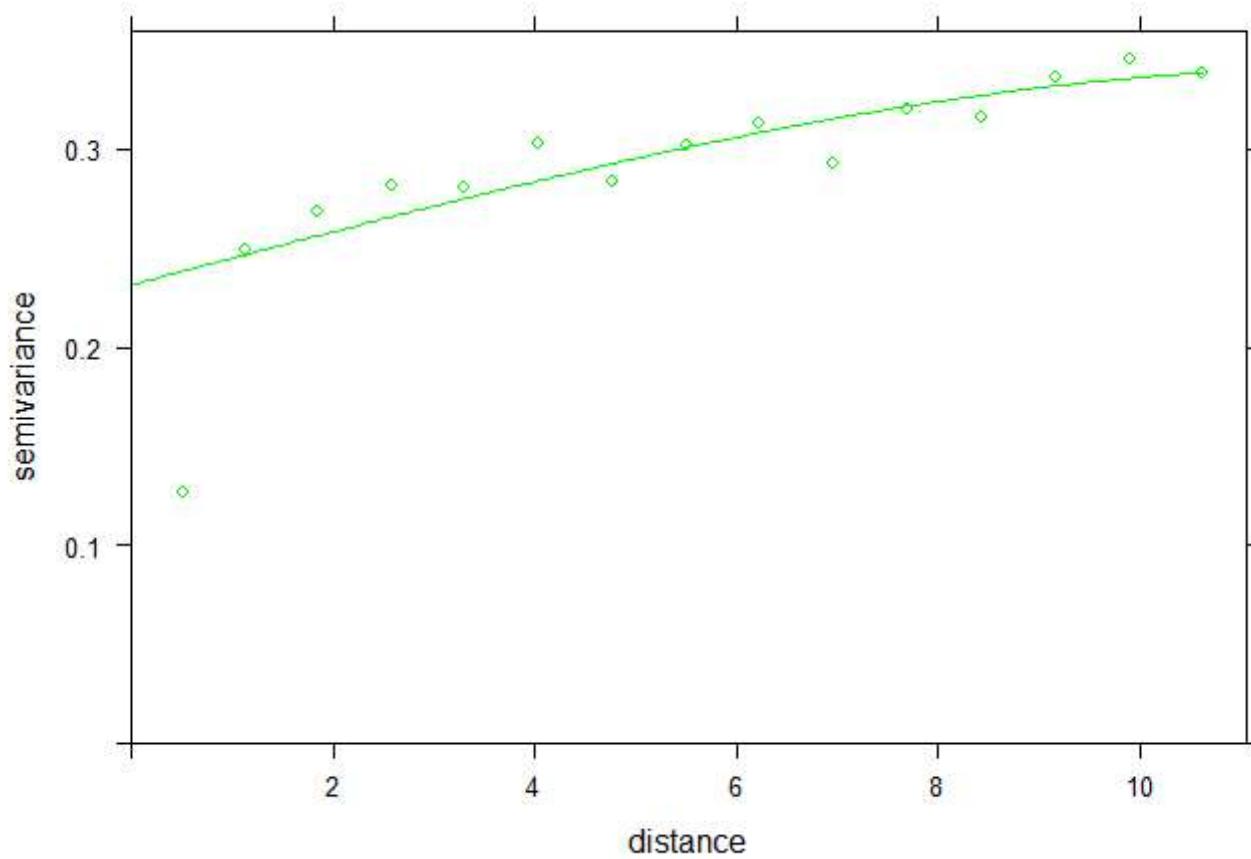


Fitting the model using different weights, we choose the OLS method as it best fits the variogram.

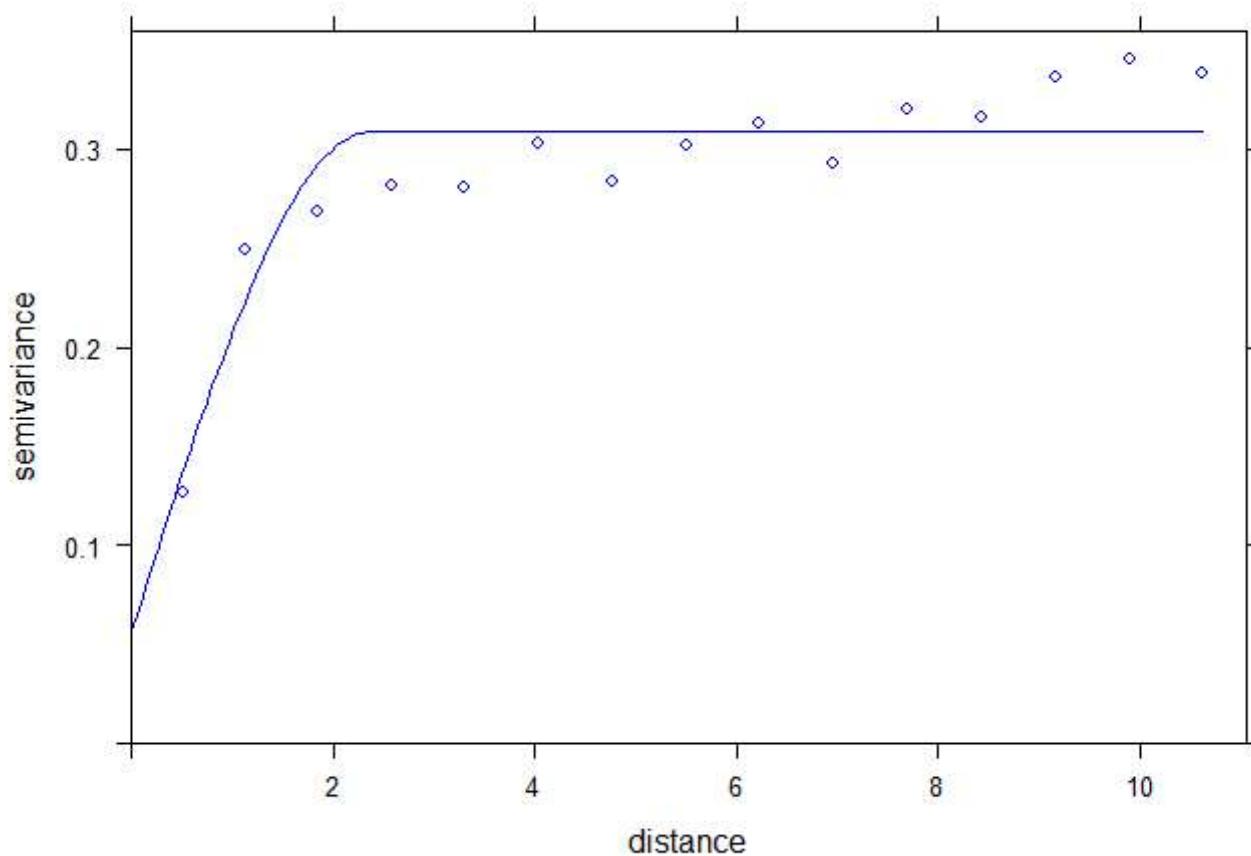
```
#fitting the model using different weights.
par(mfrow=c(2,1))
plot(vario_trend,fit.variogram(vario_trend, vgm(0.1,"Sph",10,0.32), fit.method=1),col='red')
```



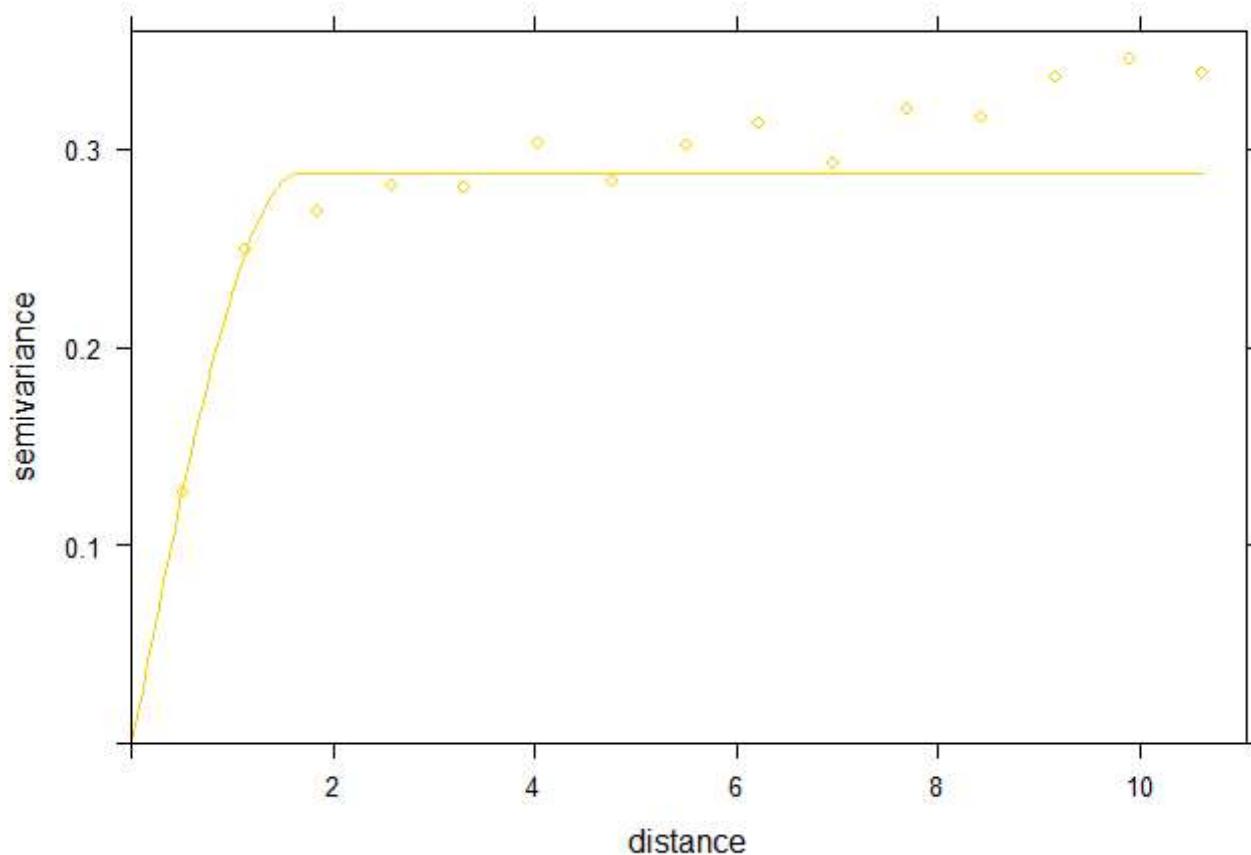
```
plot(vario_trend,fit.variogram(vario_trend, vgm(0.1,"Sph",10,0.32), fit.method=2),col='green')
```



```
plot(vario_trend,fit.variogram(vario_trend, vgm(0.1,"Sph",10,0.32), fit.method=6),col='blue')
```



```
plot(vario_trend,fit.variogram(vario_trend, vgm(0.1,"Sph",10,0.32), fit.method=7),col='gold')
```



### -Variocloud & Robust Estimator-

Plotting varigrams using both classical and robust estimator, we see that using the robust estimator results in significantly less outliers.

```
#computing variogram in east-west direction
par(mfrow=c(1,2))
varioog_ew <- variog(geo_data,direction = pi/2)
```

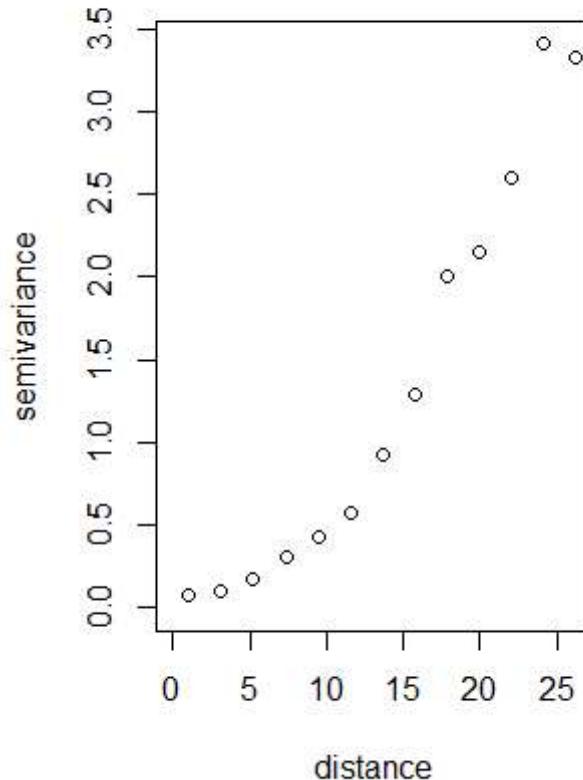
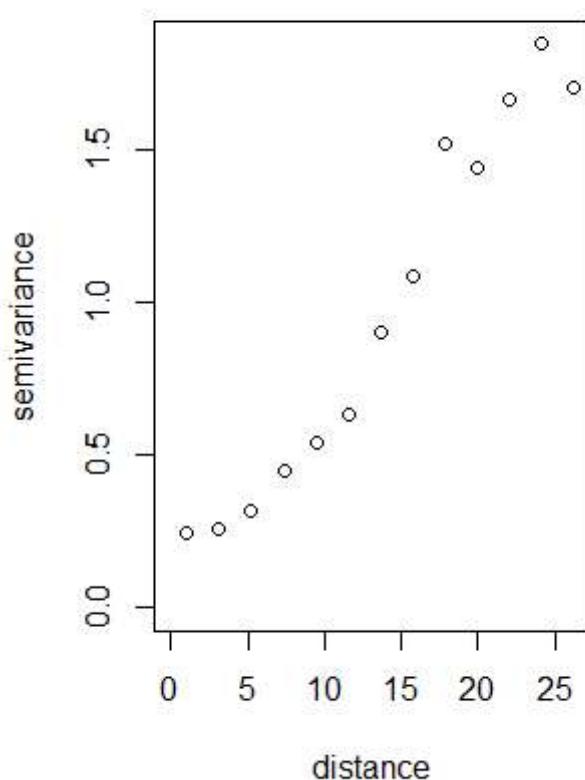
```
## variog: computing variogram for direction = 90 degrees (1.571 radians)
##           tolerance angle = 22.5 degrees (0.393 radians)
```

```
plot(varioog_ew)

#computing with robust estimator
varioog_ew_rob <- variog(geo_data,direction = pi/2,estimator.type = "modulus")
```

```
## variog: computing variogram for direction = 90 degrees (1.571 radians)
##          tolerance angle = 22.5 degrees (0.393 radians)
```

```
plot(variog_ew_rob)
```



```
#fit both variograms
par(mfrow=c(1,2))
fit_ew <- variofit(variog_ew,cov.model = "gau",ini.cov.pars = c(1.2,24),nugget = 0.2,fix.nugget = T)
```

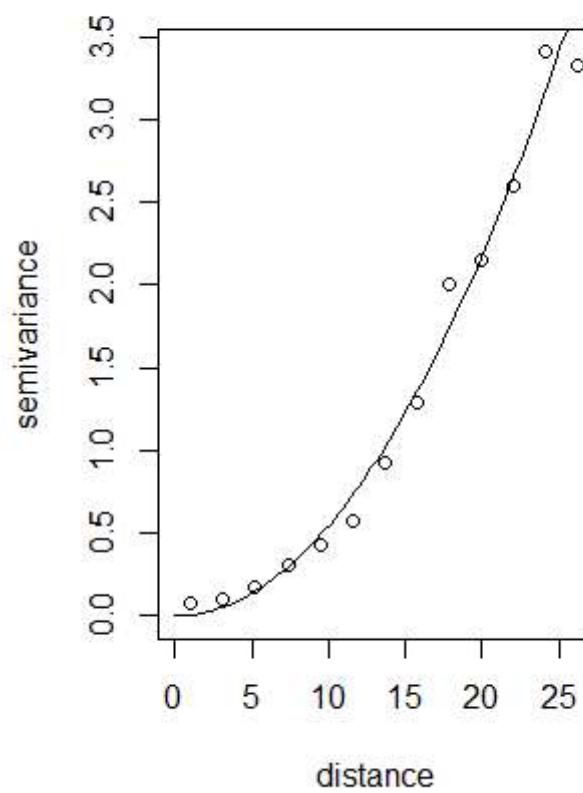
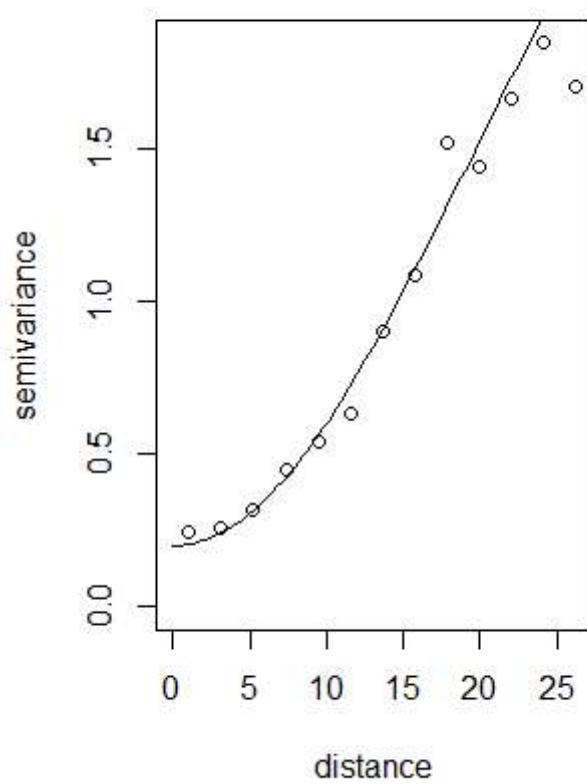
```
## variofit: covariance model used is gaussian
## variofit: weights used: npairs
## variofit: minimisation function used: optim
```

```
plot(variog_ew)
lines(fit_ew)

fit_ew_rob <- variofit(variog_ew_rob,cov.model="gau",ini.cov.pars = c(3.4,24),nugget = 0,fix.nugget = T)
```

```
## variofit: covariance model used is gaussian
## variofit: weights used: npairs
## variofit: minimisation function used: optim
```

```
plot(variog_ew_rob)
lines(fit_ew_rob)
```



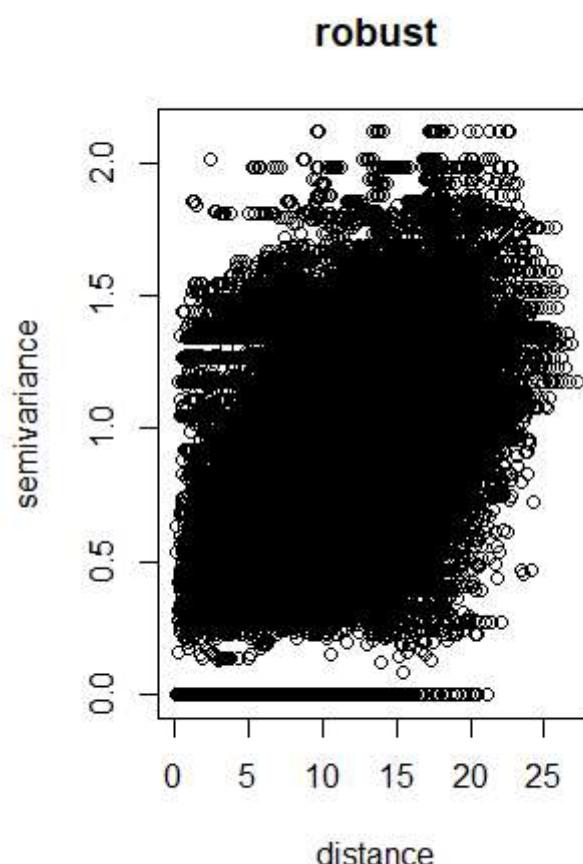
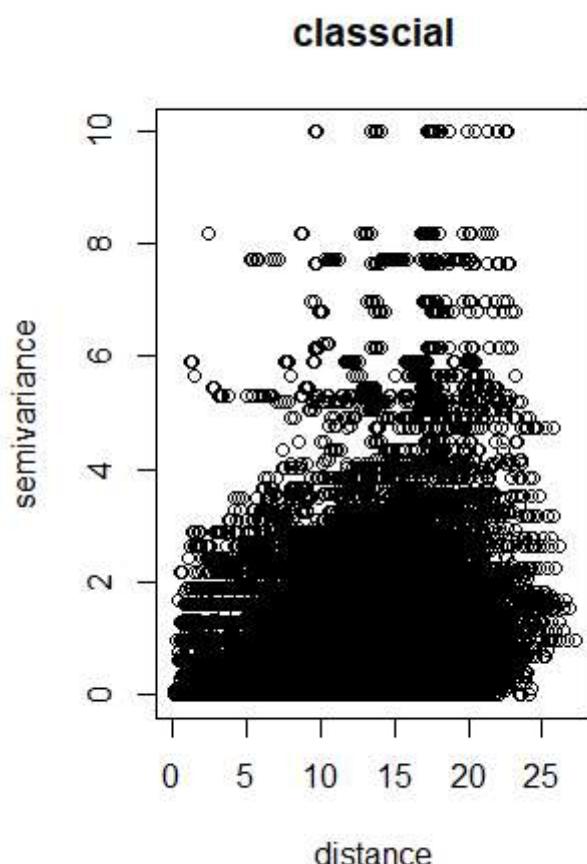
```
#compute variogram cloud
cloud <- variog(geo_data,direction = pi/2,option = "cloud")
```

```
## variog: computing variogram for direction = 90 degrees (1.571 radians)
##           tolerance angle = 22.5 degrees (0.393 radians)
```

```
cloud_rob <- variog(geo_data,direction = pi/2,estimator.type = "modulus", option = "cloud")
```

```
## variog: computing variogram for direction = 90 degrees (1.571 radians)
##           tolerance angle = 22.5 degrees (0.393 radians)
```

```
par(mfrow=c(1,2))
plot(cloud)
title(main = "classcial")
plot(cloud_rob)
title(main = "robust")
```



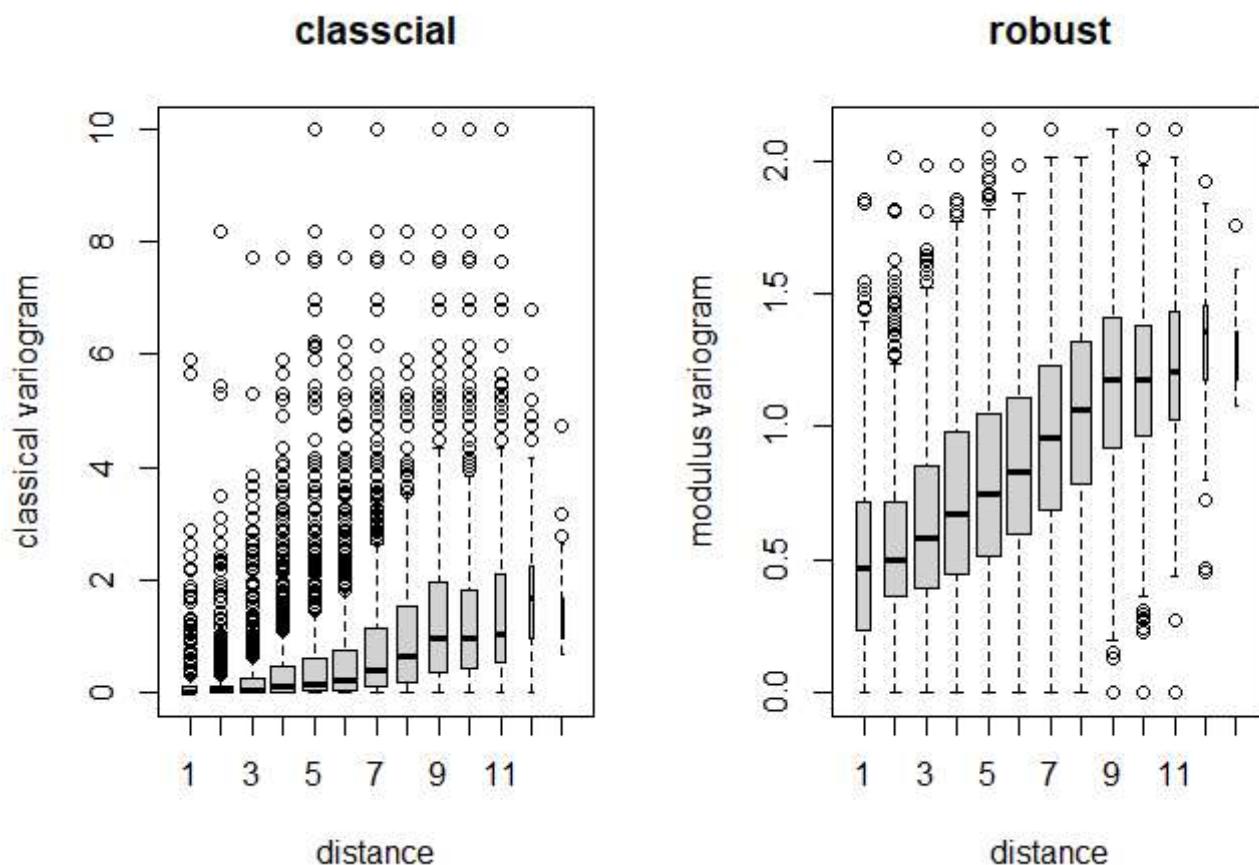
```
#compute box plot for each cloud
bar <- variog(geo_data,direction = pi/2,bin.cloud = T)
```

```
## variog: computing variogram for direction = 90 degrees (1.571 radians)
##          tolerance angle = 22.5 degrees (0.393 radians)
```

```
bar_rob <- variog(geo_data,direction = pi/2,estimator.type = "modulus", bin.cloud = T)
```

```
## variog: computing variogram for direction = 90 degrees (1.571 radians)
##          tolerance angle = 22.5 degrees (0.393 radians)
```

```
par(mfrow=c(1,2))
plot(bar,bin.cloud=T)
title(main = "classcial")
plot(bar_rob,bin.cloud=T)
title(main = "robust")
```



## Kriging

Finally, let's use the kriging method to predict the data.

### -Ordinary Kriging-

We first perform ordinary kriging.

```
#conduct ordinary kriging using gstat
#create the grid:
data_ok<-data_n[,c(1,2,3,4,6)]
names(data_ok)[c(1,2)] <- c("x","y")
x.range <- as.integer(range(data_ok[,1]))
x.range
```

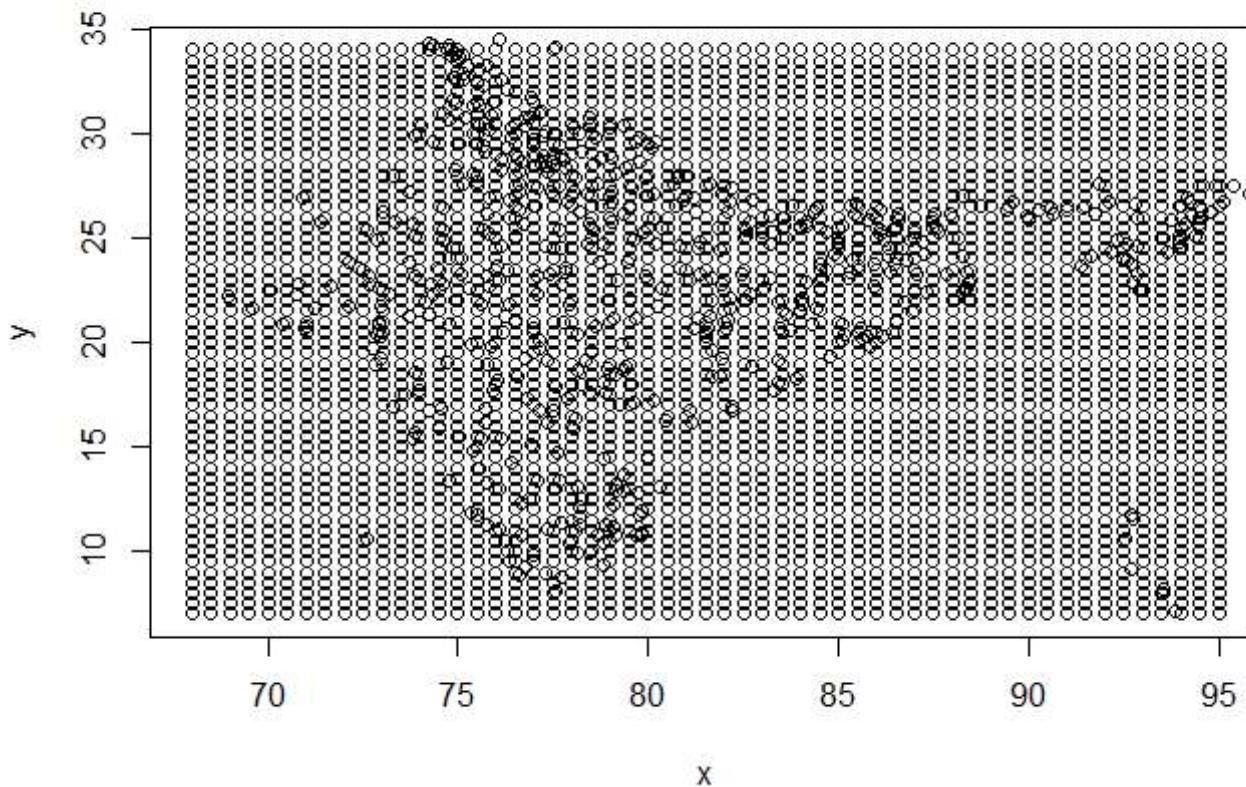
```
## [1] 68 95
```

```
y.range <- as.integer(range(data_ok[,2]))
y.range
```

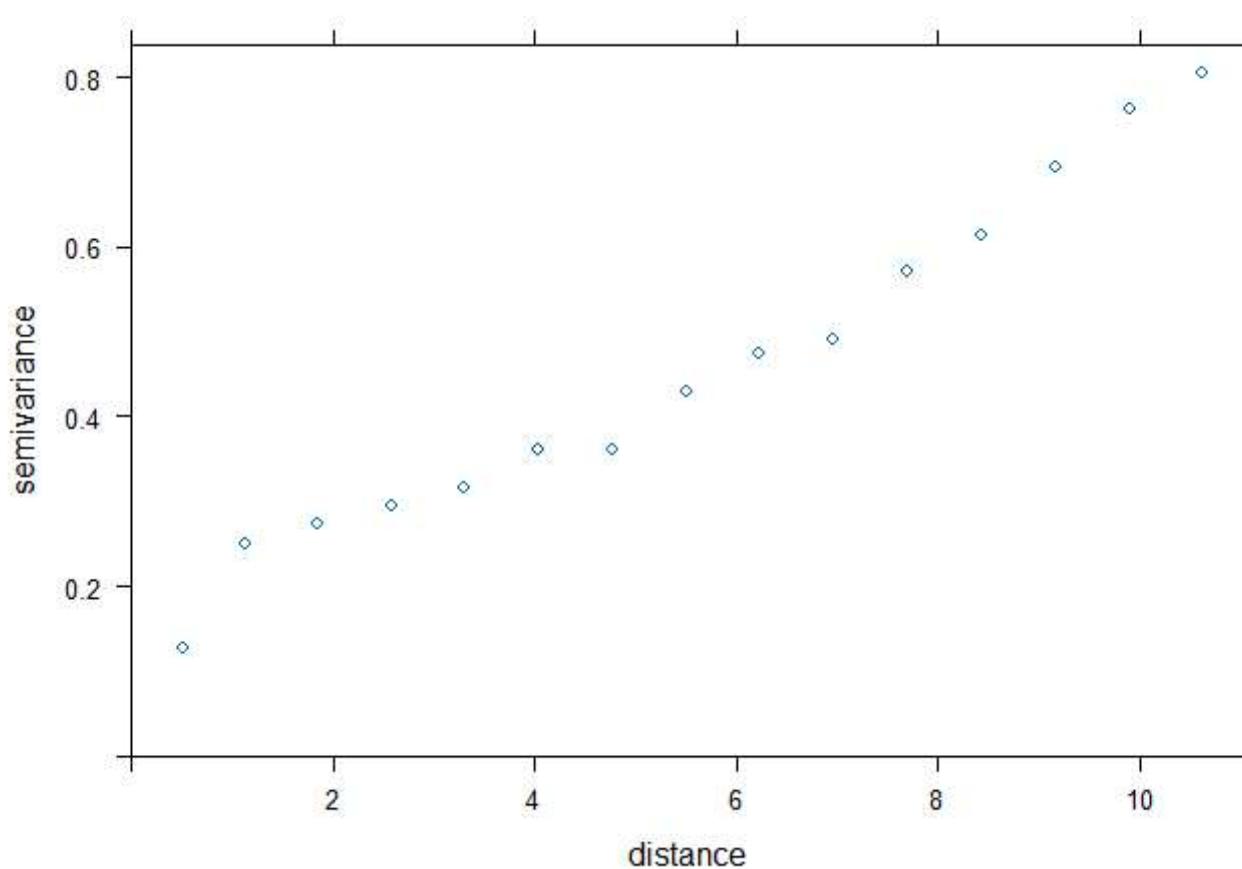
```
## [1] 7 34
```

```
grd <- expand.grid(x=seq(from=x.range[1], to=x.range[2], by=0.5),
y=seq(from=y.range[1], to=y.range[2], by=0.5))
```

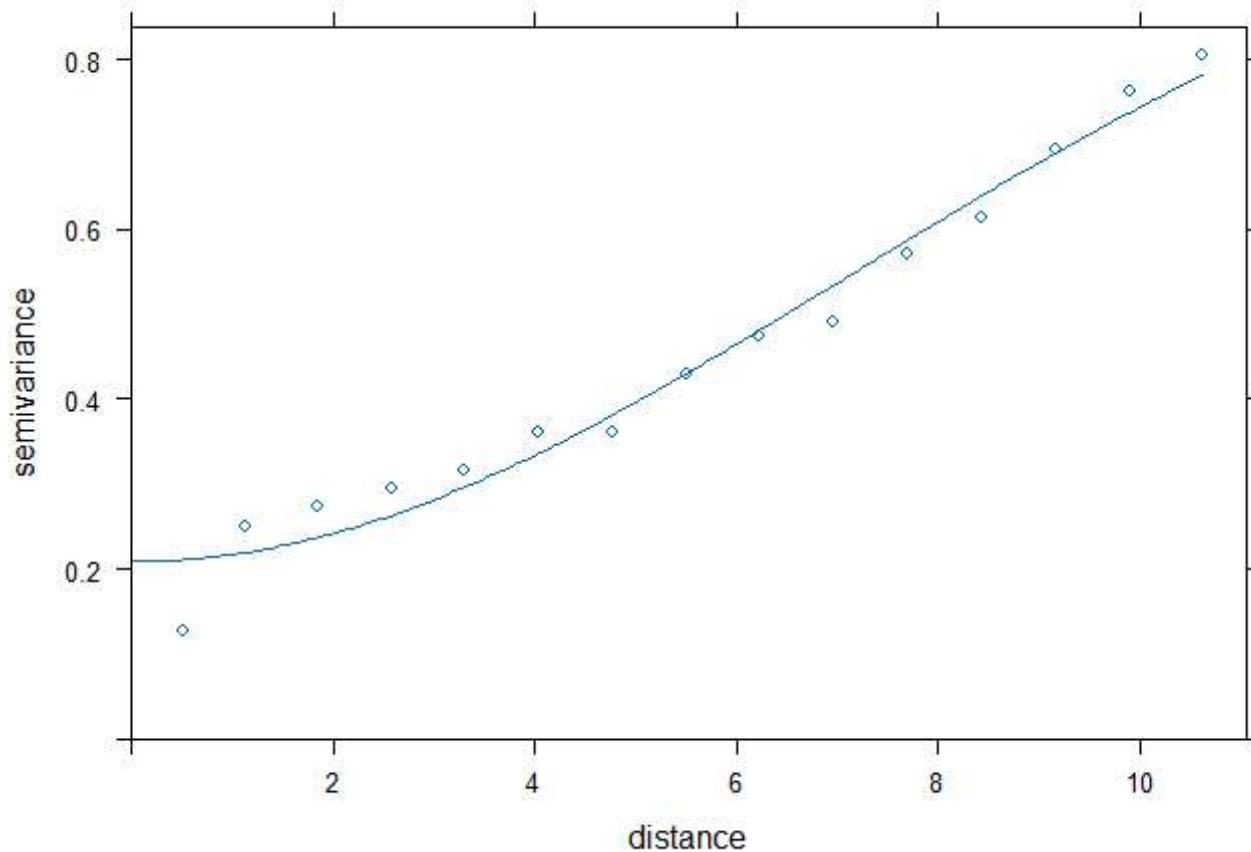
```
#Visualize the grid and the data points:  
plot(grd)  
points(data_ok)
```



```
#create gstat object and compute the sample variogram  
g <- gstat(id="log_cloud", formula = log_cloud~1, locations = ~x+y, data = data_ok)  
q <- variogram(g,cutoff=11,)  
plot(q)
```



```
#fit the variogram  
v.fit <- fit.variogram(q, vgm(0,"Gau",10,0.32), fit.method=6)  
plot(q,v.fit)
```



```
fit1$cov.pars
```

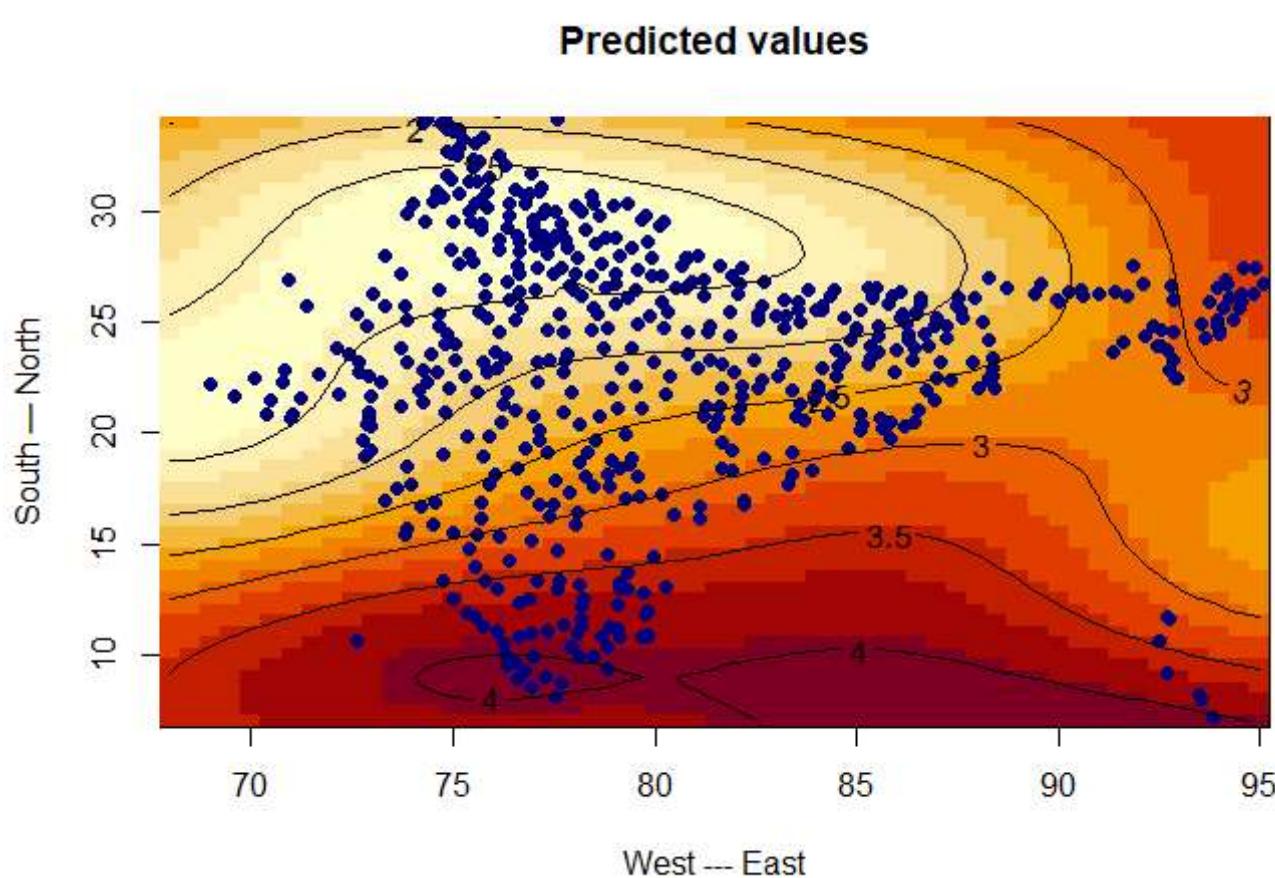
```
## [1] 4.617356 27.101523
```

```
pr_ok <- krige(id="log_cloud", log_cloud~1, locations=~x+y, model=v.fit, data=data_ok, newdata=grd)
```

```
## [using ordinary kriging]
```

```
qqq <- matrix(pr_ok$log_cloud.pred,
length(seq(from=x.range[1], to=x.range[2], by=0.5)),
length(seq(from=y.range[1], to=y.range[2], by=0.5)) )

image(seq(from=x.range[1], to=x.range[2], by=0.5),
seq(from=y.range[1], to=y.range[2], by=0.5), qqq,
xlab="West --- East", ylab="South --- North", main="Predicted values")
points(data_ok, pch=19, col='navy')
contour(seq(from=x.range[1], to=x.range[2], by=0.5),
seq(from=y.range[1], to=y.range[2], by=0.5), qqq,
add=TRUE, col="black", labcex=1)
```



## -Universal Kriging-

We will also perform universal kriging as there is a sign of spatial correlation.

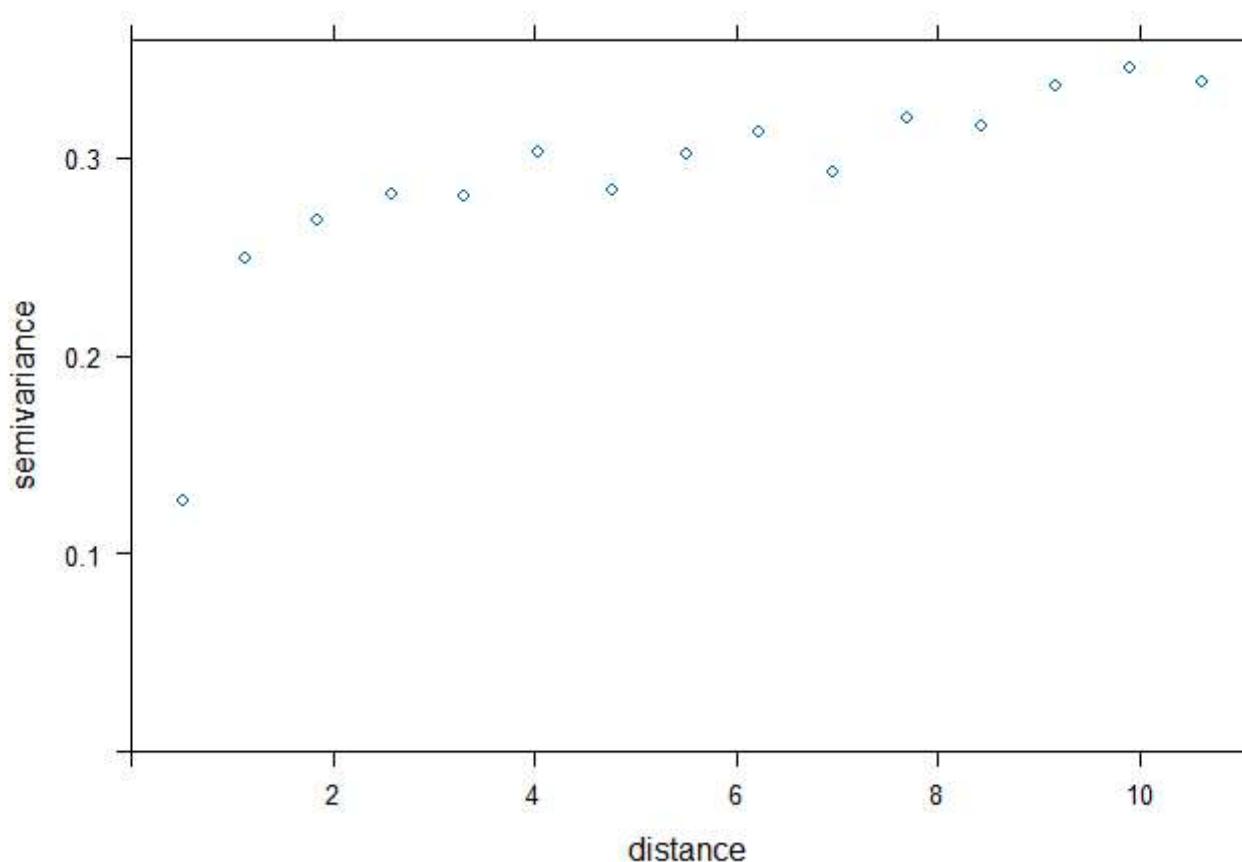
```
#conduct universal kriging using gstat  
#Create the grid:  
data_ok<-data_n[,c(1,2,3,4,6)]  
names(data_ok)[c(1,2)] <- c("x","y")  
x.range <- as.integer(range(data_ok[,1]))  
x.range
```

```
## [1] 68 95
```

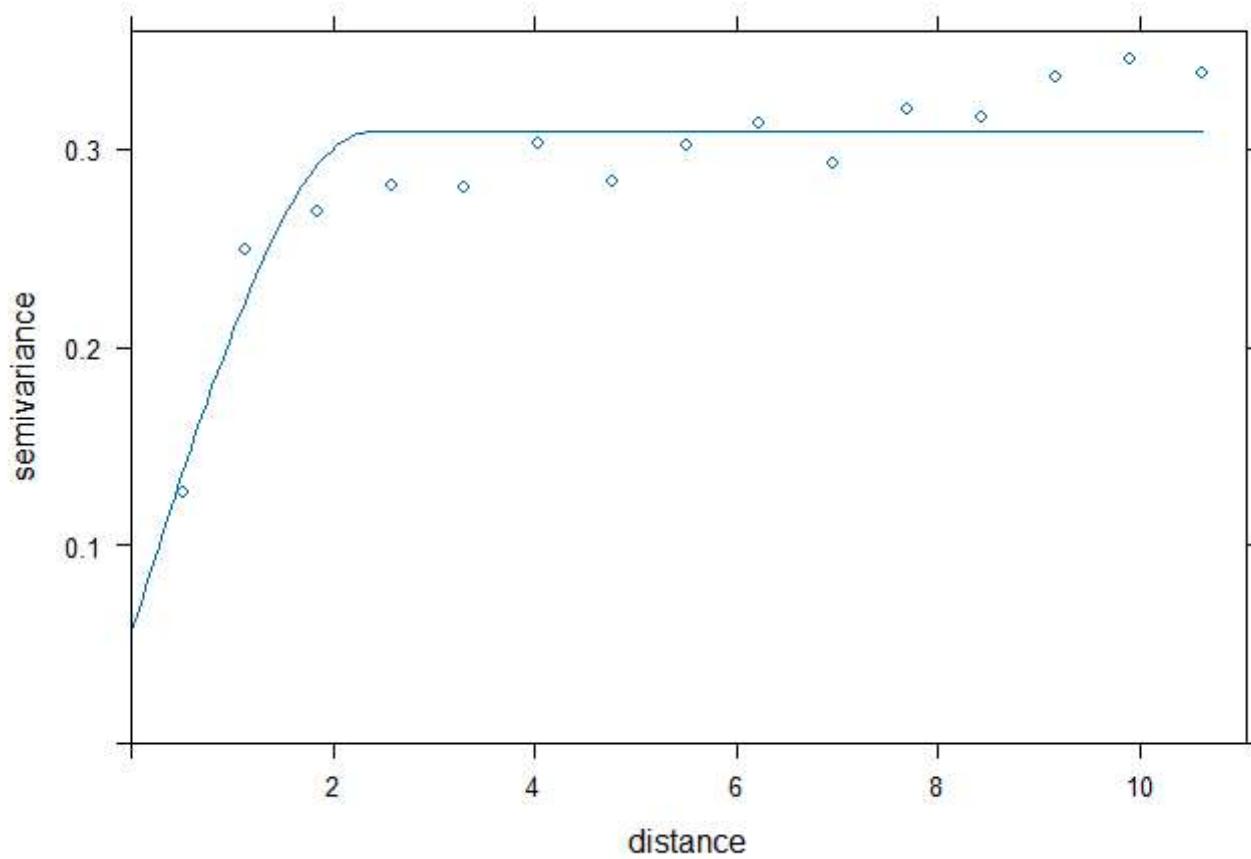
```
y.range <- as.integer(range(data_ok[,2]))  
y.range
```

```
## [1] 7 34
```

```
grd <- expand.grid(x=seq(from=x.range[1], to=x.range[2], by=0.5),  
y=seq(from=y.range[1], to=y.range[2], by=0.5))  
  
#create gstat object and compute the sample variogram  
g <- gstat(id="log_cloud", formula = log_cloud~x+y, locations = ~x+y, data = data_ok)  
q <- variogram(g,cutoff=11,)  
plot(q)
```



```
#fit the variogram based on results in previous work  
v.fit <- fit.variogram(q, vgm(0.1,"Sph",10,0.32), fit.method=6)  
plot(q,v.fit)
```



```
fit1$cov.pars
```

```
## [1] 4.617356 27.101523
```

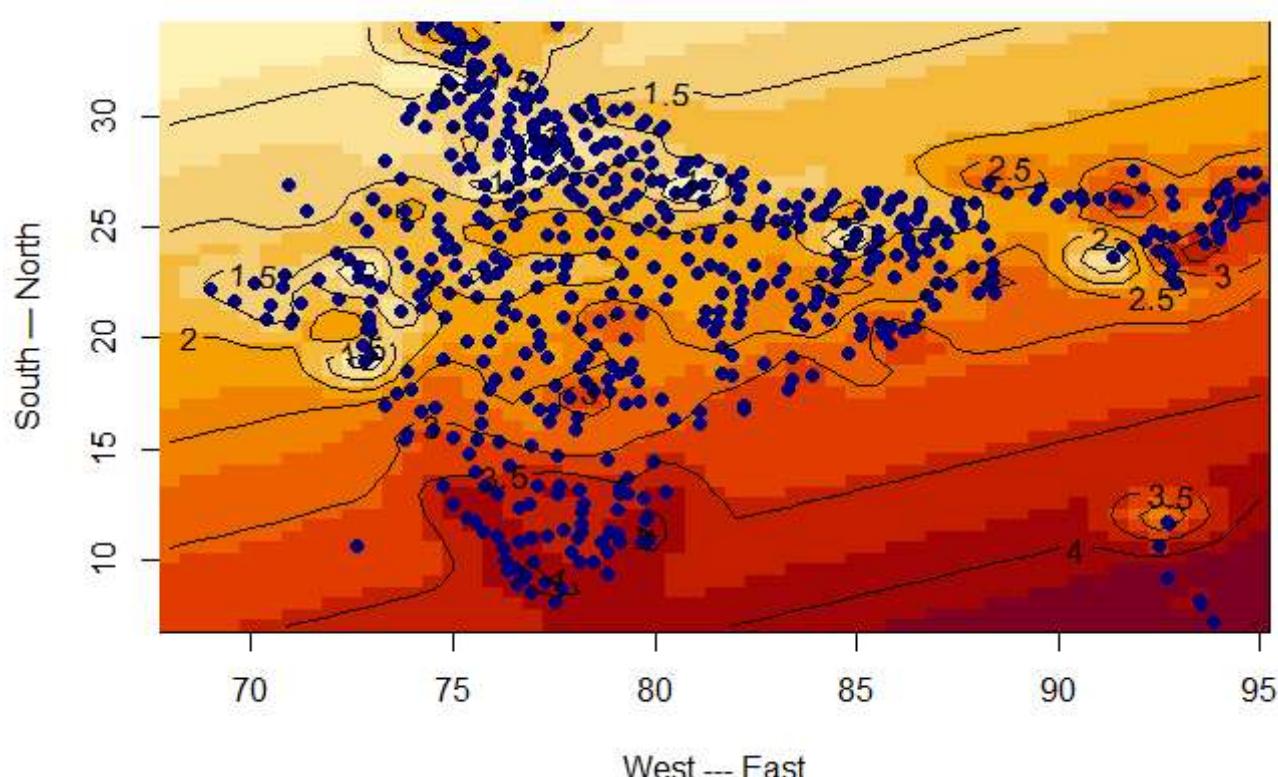
```
pr_uk <- krige(id="log_cloud", log_cloud~x+y, locations=~x+y, model=v.fit, data=data_ok, newdata=grd
```

```
## [using universal kriging]
```

```
qqq <- matrix(pr_uk$log_cloud.pred,
length(seq(from=x.range[1], to=x.range[2], by=0.5)),
length(seq(from=y.range[1], to=y.range[2], by=0.5)) )

image(seq(from=x.range[1], to=x.range[2], by=0.5),
seq(from=y.range[1], to=y.range[2], by=0.5), qqq,
xlab="West --- East", ylab="South --- North", main="Predicted values")
points(data_ok,pch=19,col='navy')
contour(seq(from=x.range[1], to=x.range[2], by=0.5),
seq(from=y.range[1], to=y.range[2], by=0.5), qqq,
add=TRUE, col="black", labcex=1)
```

### Predicted values



## -Cross Validation Between OK & UK-

From the graphs we could see universal kriging provides a better prediction of the data. Comparing the PRESS value between the two models, universal kriging has smaller PRESS value, thus we chose universal kriging for the moment.

```
#cross validation for ordinary kriging
cv_ok<- krige.cv(log_cloud~1,data=data_ok, locations=~x+y,
model=v.fit,nfold=nrow(data_ok))
```

```
#calculate the PRESS
sum(cv_ok$residual^2)/nrow(data_ok)
```

```
## [1] 0.1463145
```

```
#cross validation for universal kriging
cv_uk<- krige.cv(log_cloud~x+y,data=data_ok, locations=~x+y,
model=v.fit,nfold=nrow(data_ok))
#calculate the PRESS
PRESS_uk<-sum(cv_uk$residual^2)/nrow(data_ok)

PRESS_ok<-sum(cv_ok$residual^2)/nrow(data_ok)

PRESS_uk
```

```
## [1] 0.1305324
```

```
PRESS_ok
```

```
## [1] 0.1463145
```

```
cat("Percentage Difference is: ", (PRESS_ok-PRESS_uk)*100/PRESS_ok,"%",sep = "")
```

```
## Percentage Difference is: 10.78643%
```

## -Co-Kriging-

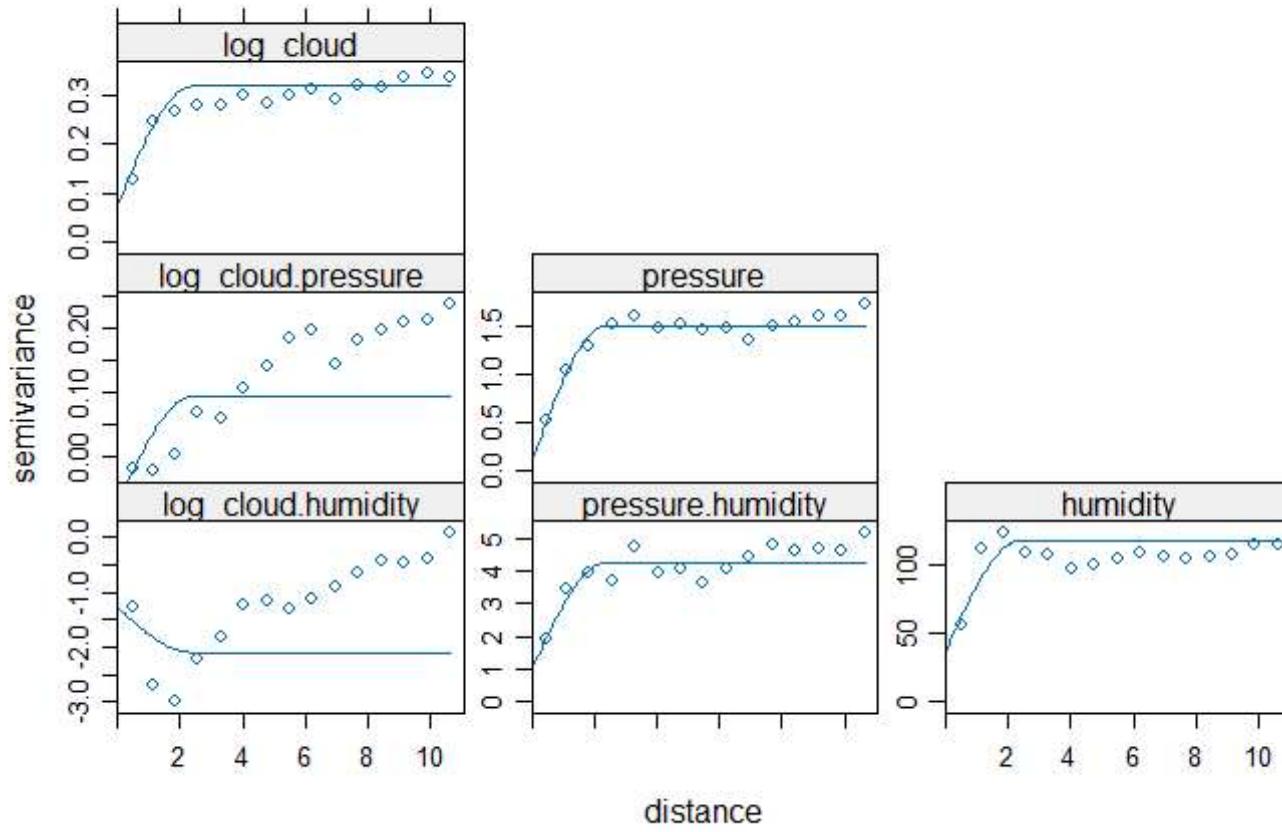
Since there is correlation between our variables, we will also perform co-kriging to see which is better.

```
g1 <- gstat(id="log_cloud", formula = log_cloud~x+y, locations = ~x+y, data = data_ok)

g1 <- gstat(g1,id="pressure", formula = avg_pressure~x+y, locations = ~x+y, data = data_ok)

g1 <- gstat(g1,id="humidity", formula = avg_humidity~x+y, locations = ~x+y, data = data_ok)

v.fit <-fit.variogram(q, vgm(0.1,"Sph",10,0.32), fit.method=6)
vm <- variogram(g1,cutoff=11)
vm.fit <-fit.lmc(vm,g1,model=v.fit)
plot(vm,vm.fit)
```



```
ck <- predict(vm.fit, grd)
```

```
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
```

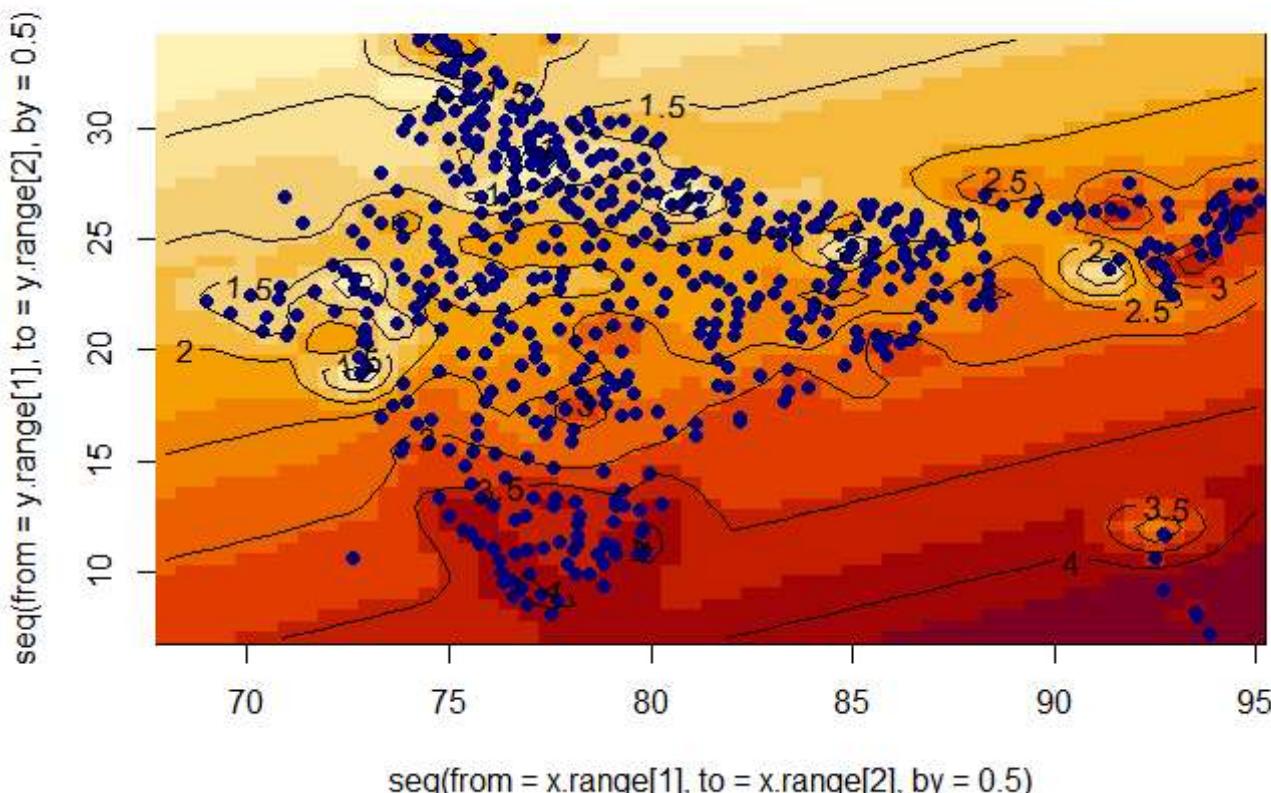
```
qq <- matrix(ck$log_cloud.pred,
length(seq(from=x.range[1], to=x.range[2], by=.5)),
length(seq(from=y.range[1], to=y.range[2], by=.5)))

image(seq(from=x.range[1], to=x.range[2], by=.5),
      seq(from=y.range[1], to=y.range[2], by=.5), qqq)
points(data_ok, pch=19, col="navy")

range(ck$log_cloud.pred)
```

```
## [1] 0.2654088 4.4714324
```

```
contour(seq(from=x.range[1], to=x.range[2], by=.5),
seq(from=y.range[1], to=y.range[2], by=.5), qqq, add=TRUE, col="black",
labcex=1)
```



## -Cross Validation Between UK & CK-

Using cross-validation, we see that universal kriging has a smaller PRESS value

```
cv_ck <- gstat.cv(vm.fit)
```































```
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
## [using universal cokriging]
## Linear Model of Coregionalization found. Good.
```

```
PRESS_ck<-sum(cv_ck$residual^2)/nrow(data_ok)
PRESS_ck
```

```
## [1] 0.1354802
```

```
PRESS_uk
```

```
## [1] 0.1305324
```

## Conclusion

Therefore, we recommend still using UNIVERSAL KRIGING to predict the data as it has a smaller PRESS value.