Ozone Concentraion in California 3/12/2020

Josh Park

3/13/2020

Introduction

Ground ozone rate is affected by the variety of reasons. Ozone can be synthesized by the chemical reaction between between oxides of nitrogen and volatile organic compounds. In the presence of sunlight, pollutants that emitted by the cars, factories, and other sources can be converted into ozone as well. I wanted to know the correlation of metropolis on ozone quality. Since I know that the California is a place full of sunlight and human being, I assume that this place will be full of ground ozone. I got the California ozone data from US Environmental Protection Agency. In the EPA, users can find the variety of pollutants such as CO, PO2, NO2, and Ozone. It also allows us to select the year and area that we want to see. So, it was easy for me to get the most recent data that I needed.

```
library(geoR)
```

```
## --------------------------------------------------------------
##  Analysis of Geostatistical Data
##  For an Introduction to geoR go to http://www.leg.ufpr.br/geoR
##  geoR version 1.7-5.2.2 (built on 2016-05-02) is now loaded
## --------------------------------------------------------------
```

```
library(gstat)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
library(ggmap)
```

```
## Loading required package: ggplot2
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## Google's Terms of Service: https://cloud.google.com/maps-platform/terms/.
```

```
## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(maps)
library(dplyr)
```

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

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

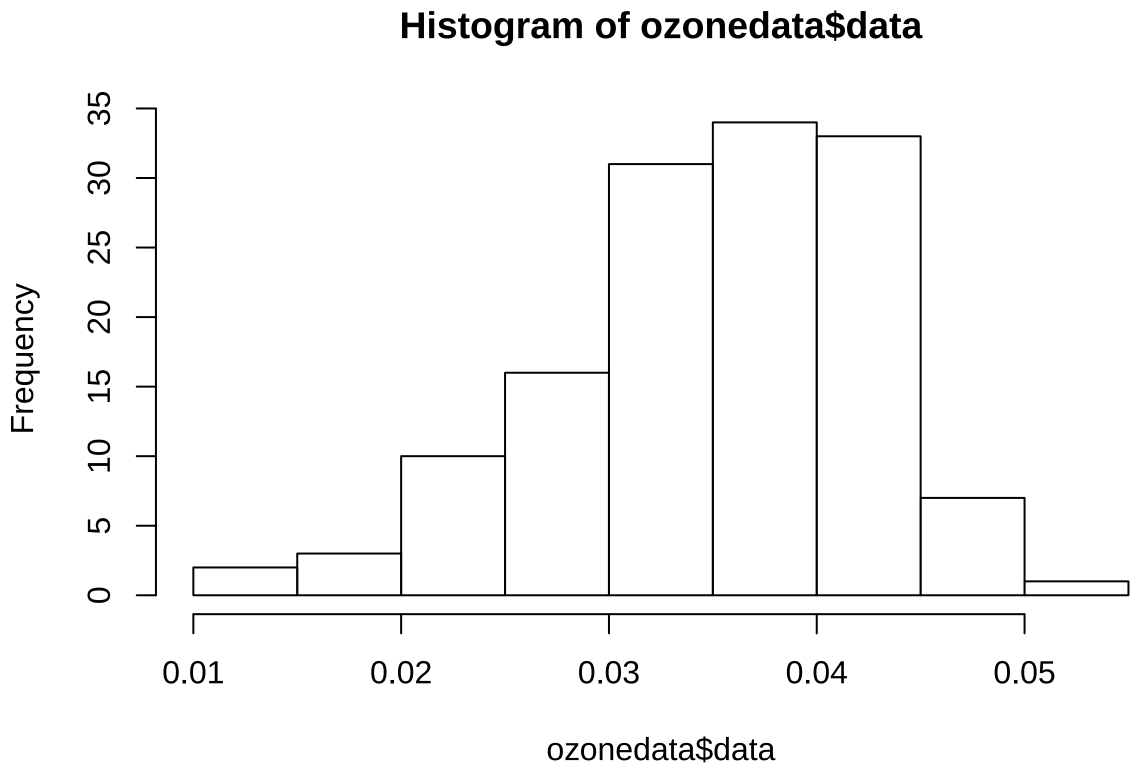
```
##
##     filter, lag
```

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

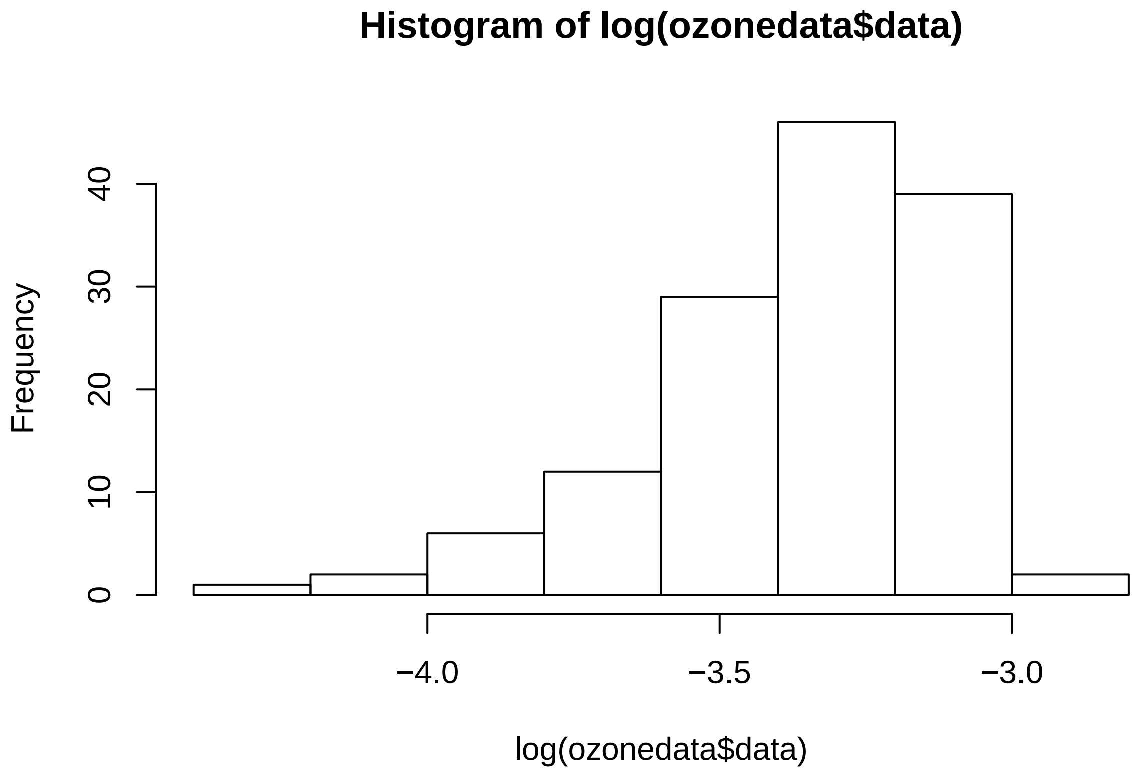
```
ozone <- read.csv("~/Dropbox/School/UCLA/stat c173/Final Project/ozone.csv", header = TRUE)
ozone$Date <- as.Date(ozone$Date, "%m/%d/%Y")
ozone_oneday <- ozone %>% filter(Date == "2020-03-12")
ozonedata <- ozone_oneday[,c(20,19,5)]
names(ozonedata) <- c("x","y","data")

# EDA
summary(ozonedata)
```

```
##        x                 y              data
##  Min.   :-124.2   Min.   :32.63   Min.   :0.01300
##  1st Qu.:-121.6   1st Qu.:34.24   1st Qu.:0.03100
##  Median :-120.1   Median :36.32   Median :0.03700
##  Mean   :-119.8   Mean   :36.23   Mean   :0.03555
##  3rd Qu.:-118.1   3rd Qu.:37.95   3rd Qu.:0.04100
##  Max.   :-114.6   Max.   :41.73   Max.   :0.05100
```

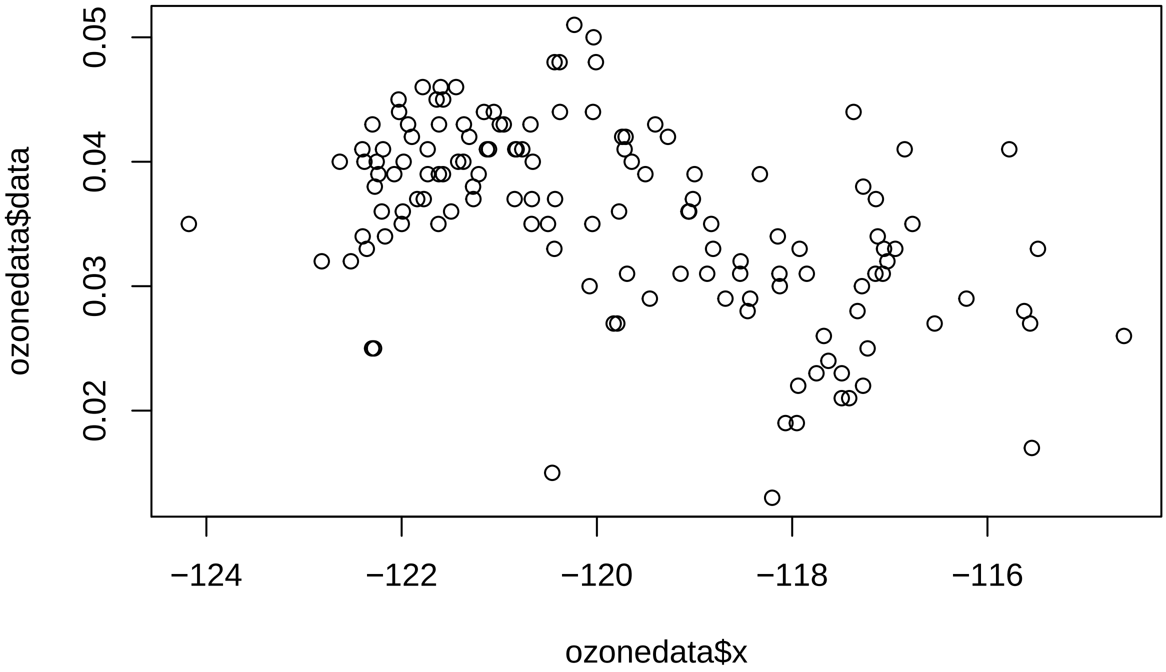
```
hist(ozonedata$data)
```



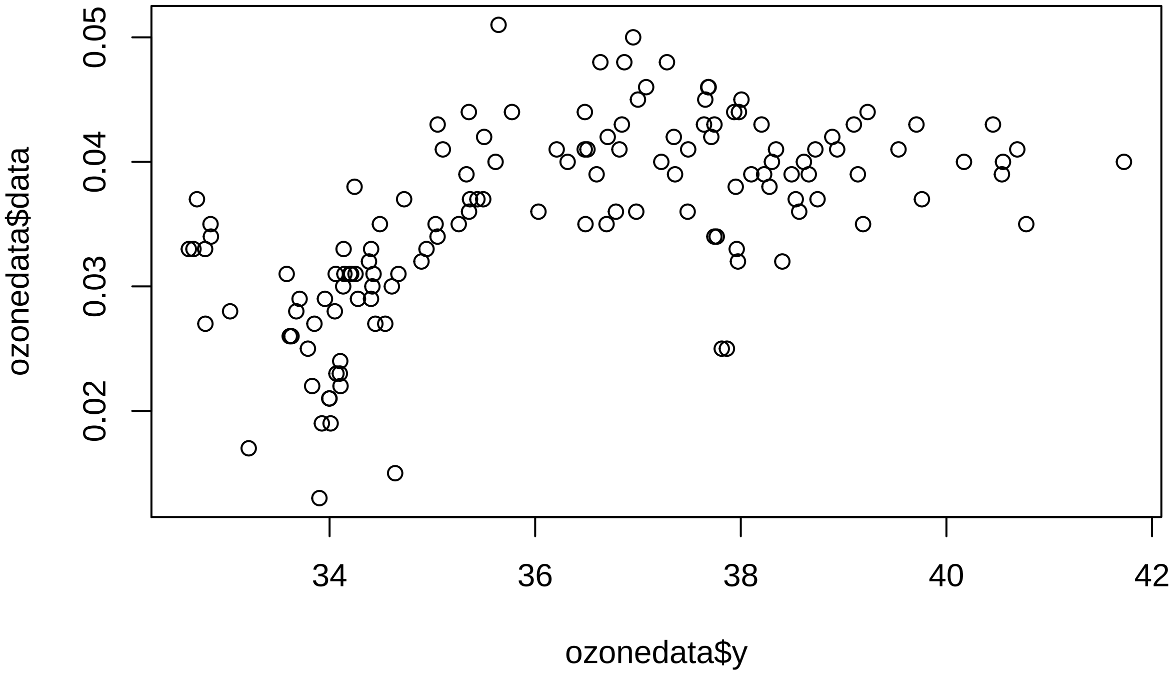
```
hist(log(ozonedata$data))
```



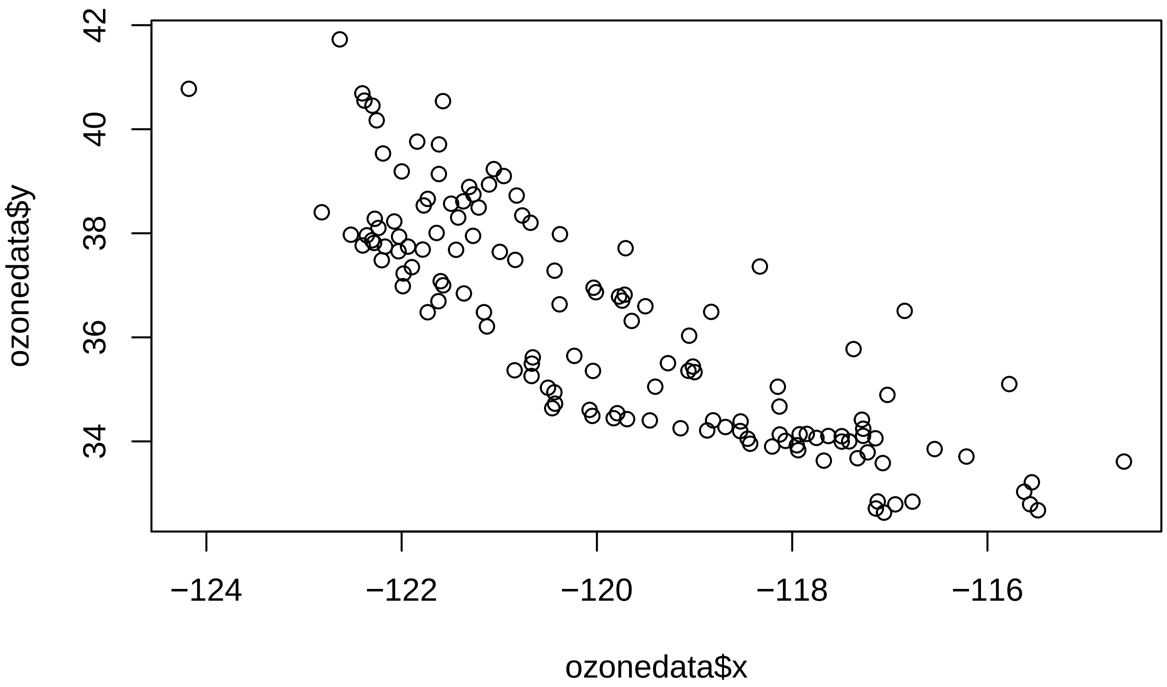
```r
plot(ozonedata$x, ozonedata$data)
```



```r
plot(ozonedata$y, ozonedata$data)
```



```
plot(ozonedata$x, ozonedata$y)
```



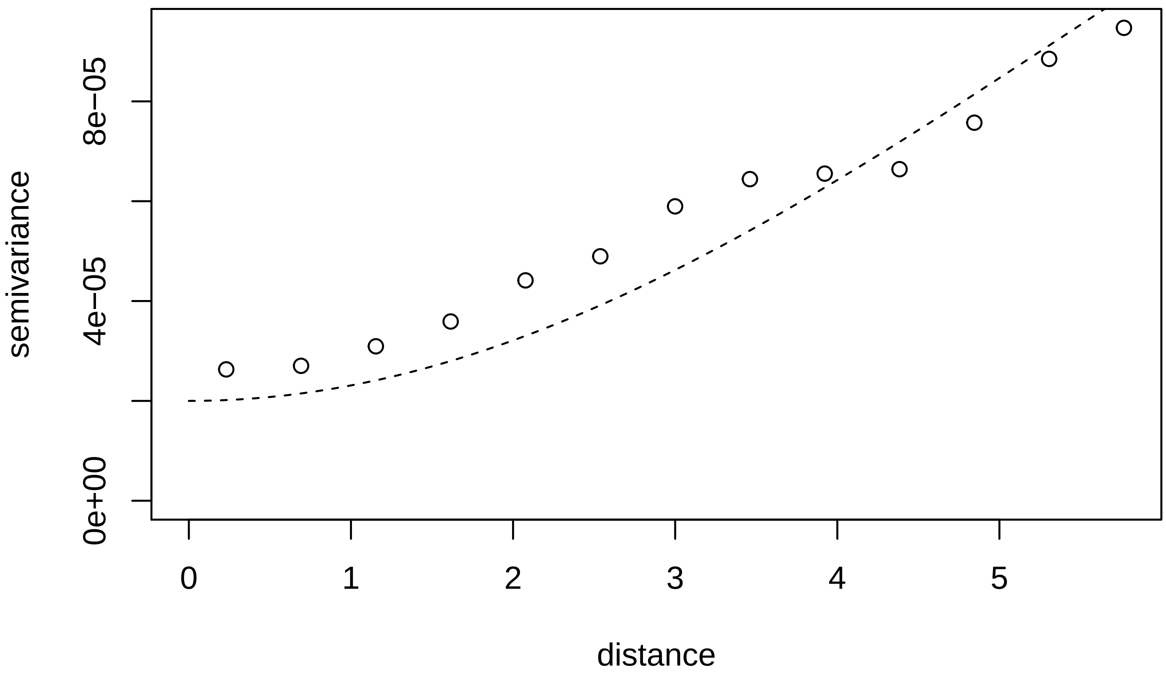
- I used daily max 8 hour ozone concentration as my dependent variable.
- In California, there are 137 locations for detecting air pollutants.
- On the middle of each coordinates, ozone concentration relatively high values like a spherical model.
- Eastern and southern California seem to have lower concentration rate.
- By watching frequency of ozone concentration, it almost lies on normal distribution which means I need no transformation to normalize it.

Crossvalidation

```
# Geodata
# Calculating sample Variogram
g_ozonedata <- as.geodata(ozonedata)
ozonedata_variogram <- variog(g_ozonedata, max.dist = 6)
```

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

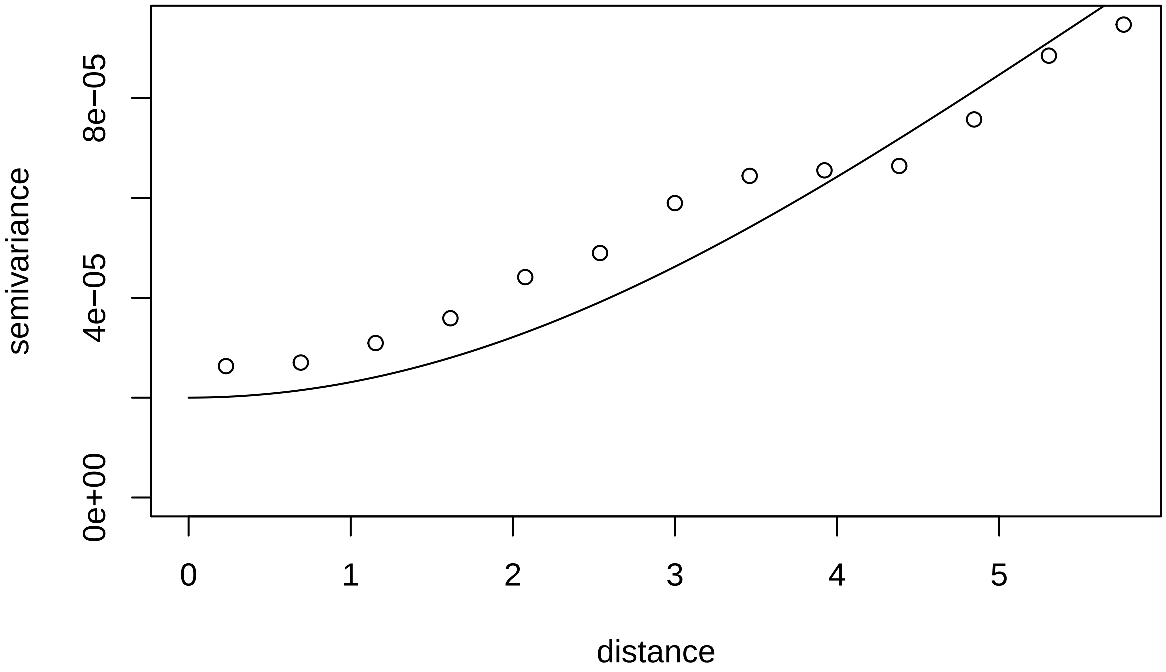
```
plot(ozonedata_variogram)
lines.variomodel(cov.model="gau", cov.pars=c(0.0002,8), nug=0.00002, max.dist=6, lty=2)
```



```
# Fitting sample Variogram
vfit5 <- variofit(ozonedata_variogram, cov.model="gau", ini.cov.pars=c(0.0002,8), fix.nugget=FALSE, nugg
```

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

```
plot(ozonedata_variogram)
lines(vfit5)
```



```
#Perform cross validation for exponential model:
x_val1 <- xvalid(g_ozonedata, model=vfit5)
```

```
## xvalid: number of data locations       = 137
## xvalid: number of validation locations = 137
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1
## xvalid: end of cross-validation
```

```
#Or re-estimating the variogram after we omit each data point:
x_val2 <- xvalid(g_ozonedata, model=vfit5, reest=TRUE, variog.obj=ozonedata_variogram)
```

```
## xvalid: number of data locations       = 137
## xvalid: number of validation locations = 137
## xvalid: performing cross-validation at location ... 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1
## xvalid: end of cross-validation
```

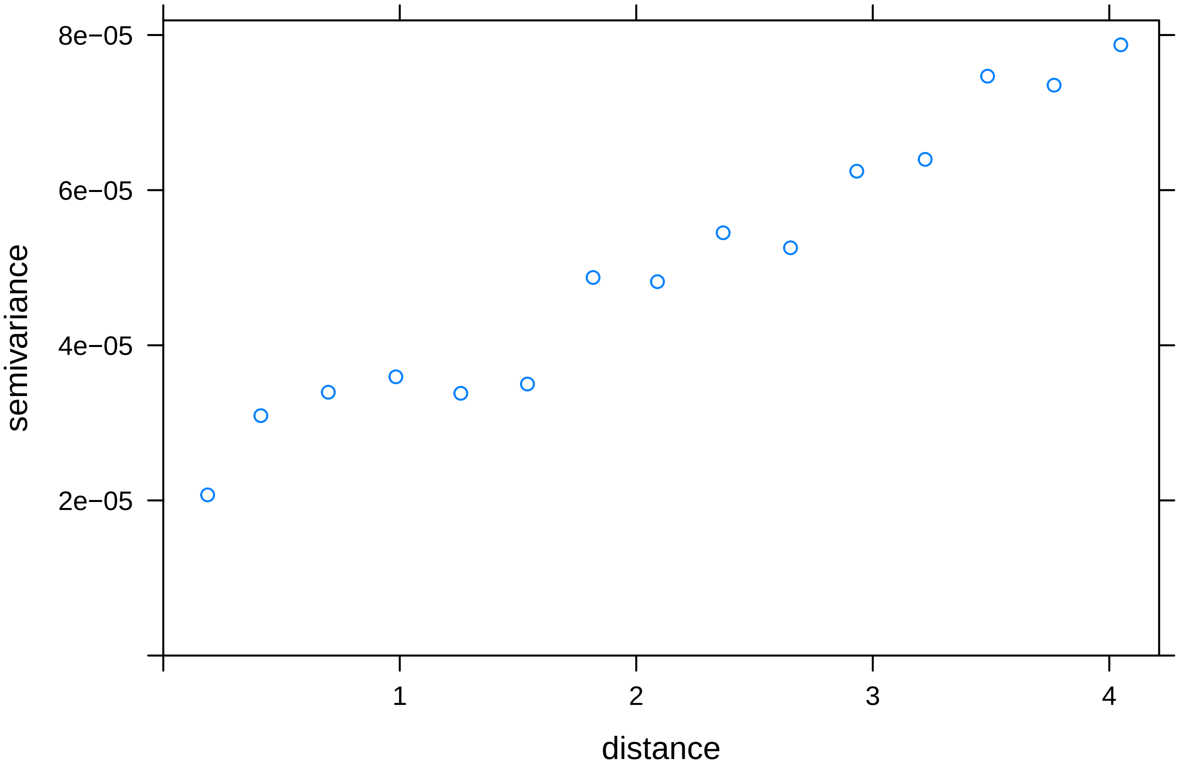
```
#Compute the prediction sum of squares:
dif <- ozonedata$data - x_val1$predicted
PRESS1 <- sum(dif^2)
PRESS1
```

```
## [1] 0.004169637
```

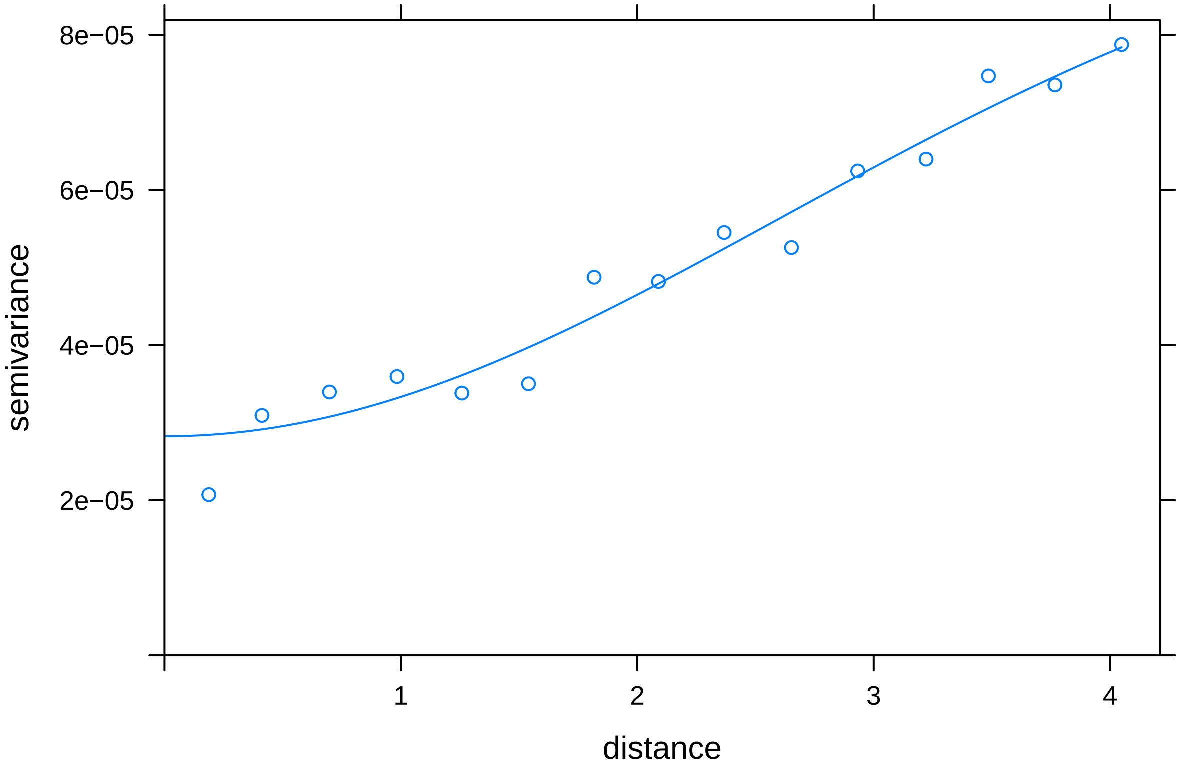
```
dif <- ozonedata$data - x_val2$predicted
PRESS2 <- sum(dif^2)
PRESS2
```

```
## [1] 0.004169637
```

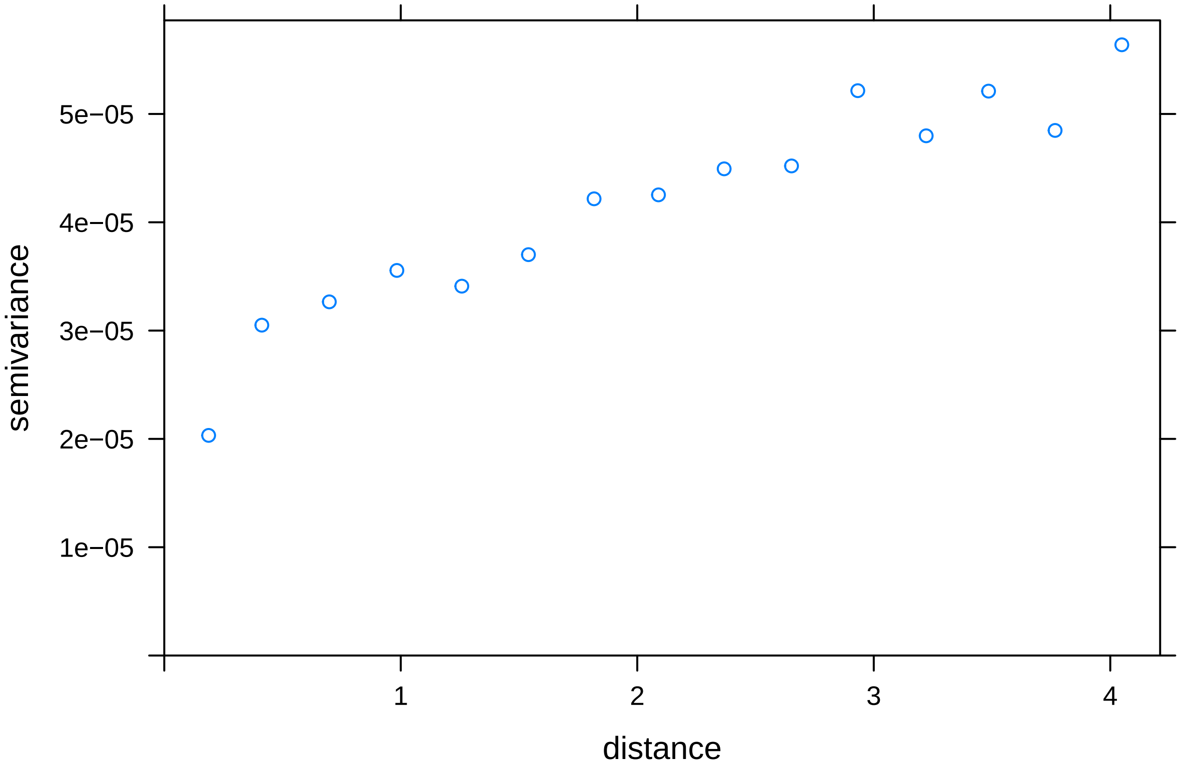
```
# Gstat
# Divide into train and test data
ozone_train <- ozonedata[c(1:120),]
ozone_test <- ozonedata[c(121:137),]

gstat_ozone_o <- gstat(id="data", formula = data~1, locations = ~x+y, data = ozone_train)
gstat_ozone_u <- gstat(id="data", formula = data~x+y, locations = ~x+y, data = ozone_train)

plot(variogram(gstat_ozone_o))
```



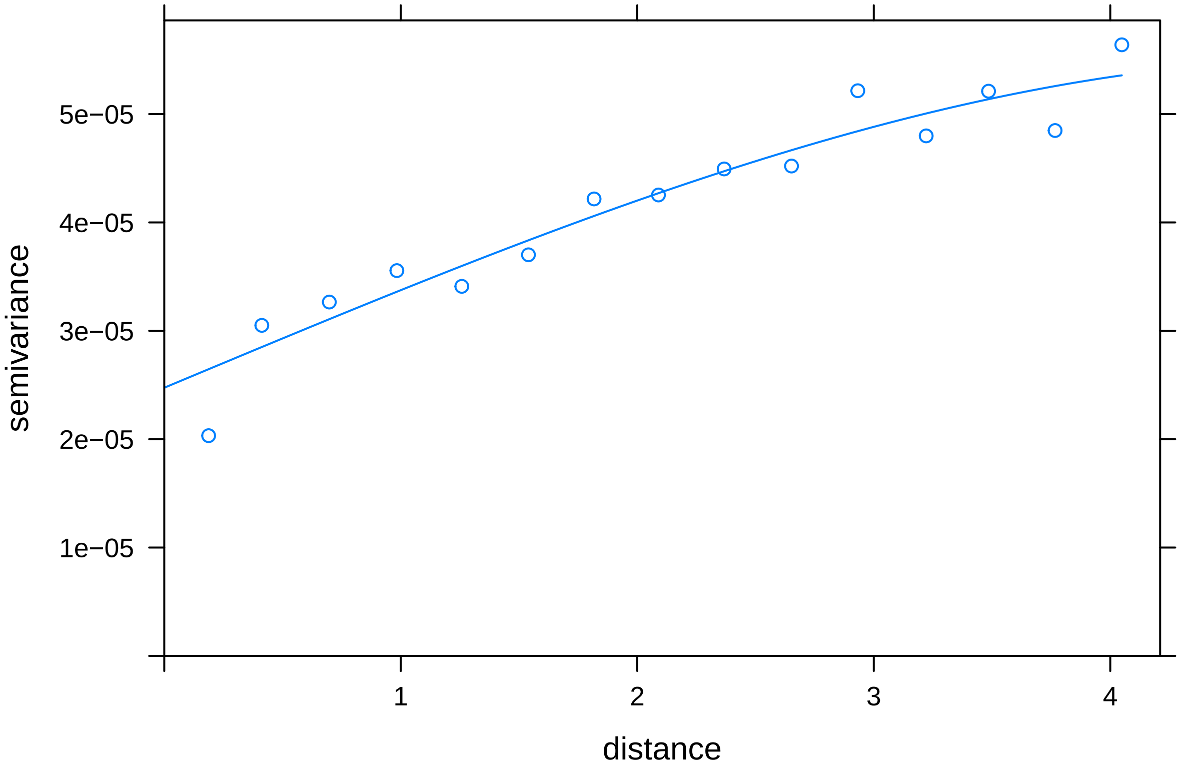
```
v.fit4 <- fit.variogram(variogram(gstat_ozone_o), vgm(0.0015,"Gau",6,0.1), fit.method=1)
plot(variogram(gstat_ozone_o), v.fit4)
```



```r
plot(variogram(gstat_ozone_u))
```



```
v.fit6 <- fit.variogram(variogram(gstat_ozone_u), vgm(0.002,"Sph",6,0.1), fit.method=1)
plot(variogram(gstat_ozone_u), v.fit6)
```



```
part_valid_pr1 <- krige(id="data", data~1, locations=~x+y, model=v.fit4, data=ozone_train, newdata=ozon
```

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

```
difference1 <- ozone_test$data - part_valid_pr1$data.pred
summary(difference1)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -7.242e-03 -1.983e-03  2.485e-05 -2.698e-04  1.403e-03  3.857e-03
```

```
press1 <- sum(difference1^2)
press1
```

```
## [1] 0.0001260581
```

```
part_valid_pr1 <- krige(id="data", data~x+y, locations=~x+y, model=v.fit6, data=ozone_train, newdata=oz
```

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

```
difference1 <- ozone_test$data - part_valid_pr1$data.pred
summary(difference1)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.0059640 -0.0016006 -0.0003429 -0.0006057  0.0015170  0.0031601
```

```
press2 <- sum(difference1^2)
press2
```

```
## [1] 0.0001080653
```

- PRESS is a criteria that calculates sum of prediction errors.

- I divided the whole data into two parts and calculated the PRESS.
- By comparing all the geodata and gstat models, gstat method with universal kriging on shperical model had lowest PRESS value.

Kriging

```
# Construct grid points for further prediction
x.range <- as.integer(range(ozonedata[,1]))
y.range <- as.integer(range(ozonedata[,2]))
x=seq(from=x.range[1], to=x.range[2], by=0.1)
y=seq(from=y.range[1], to=y.range[2], by=0.1)
grd <- expand.grid(x=x,y=y)

m <- vgm(2.474746e-05, "Sph", 4.986957, 3.033293e-05)

# Ordinary Kriging
okriging <- krige(id="data", formula = data~1, data=ozonedata, newdata=grd, model = m, locations=~x+y)
```

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

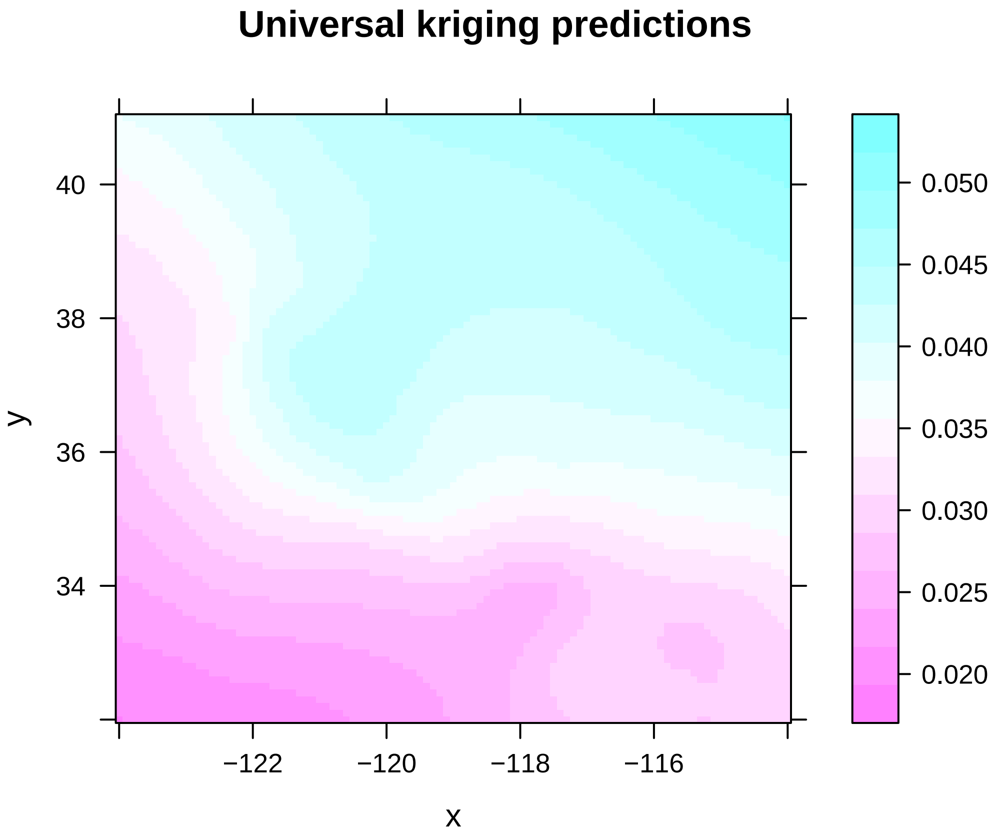
```
# Universal Kriging
ukriging <- krige(id="data", formula = data~x+y, data=ozonedata, newdata=grd, model = m, locations=~x+y
```

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

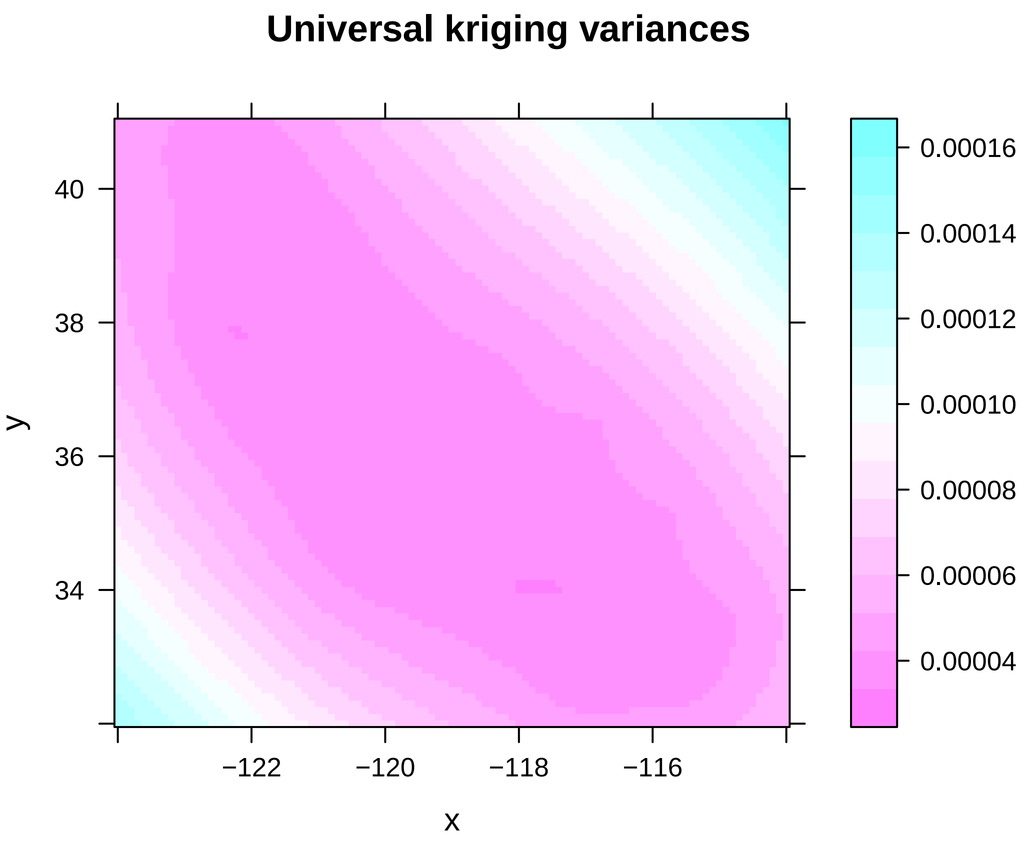
```
#Compare the three methods:
pred <- cbind(okriging$data.pred, ukriging$data.pred)
pred.var <- cbind(okriging$data.var, ukriging$data.var)
```

Mapping

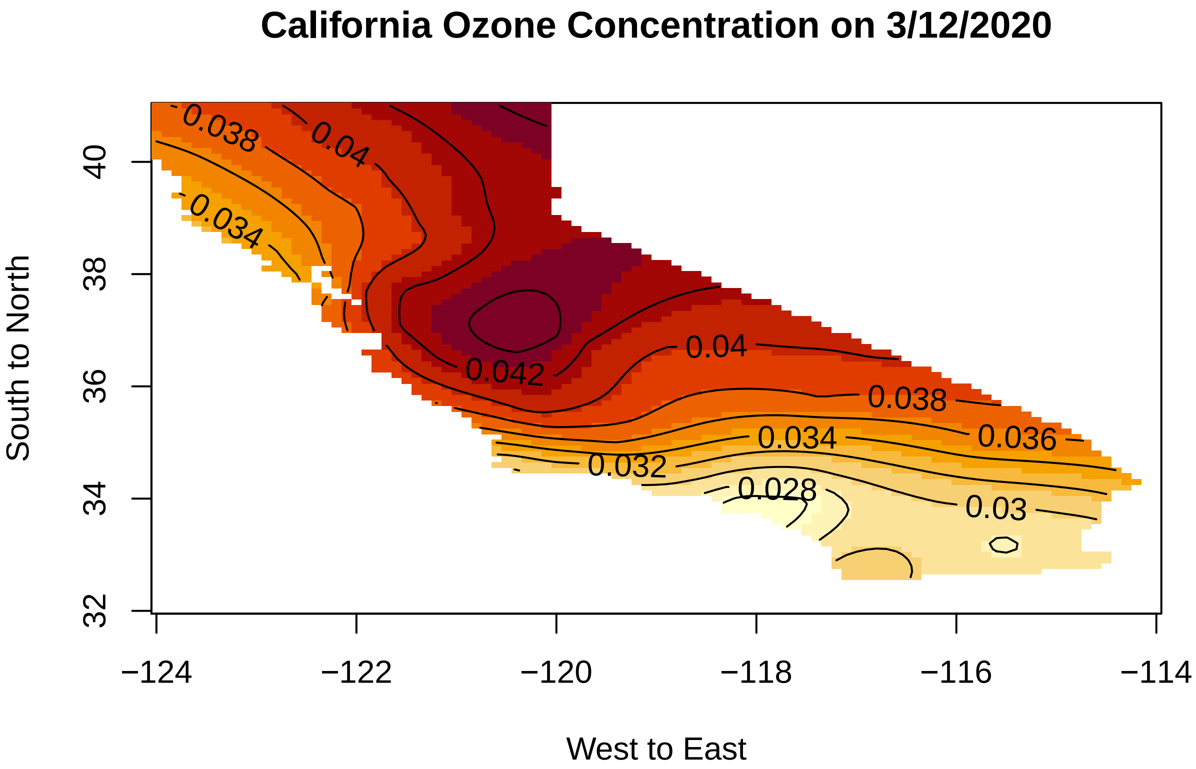
```
#Load the lattice package:
library(lattice)
#Construct a raster map using the kriged values:
levelplot(ukriging$data.pred~x+y, ukriging, aspect ="iso", main="Universal kriging predictions")
```



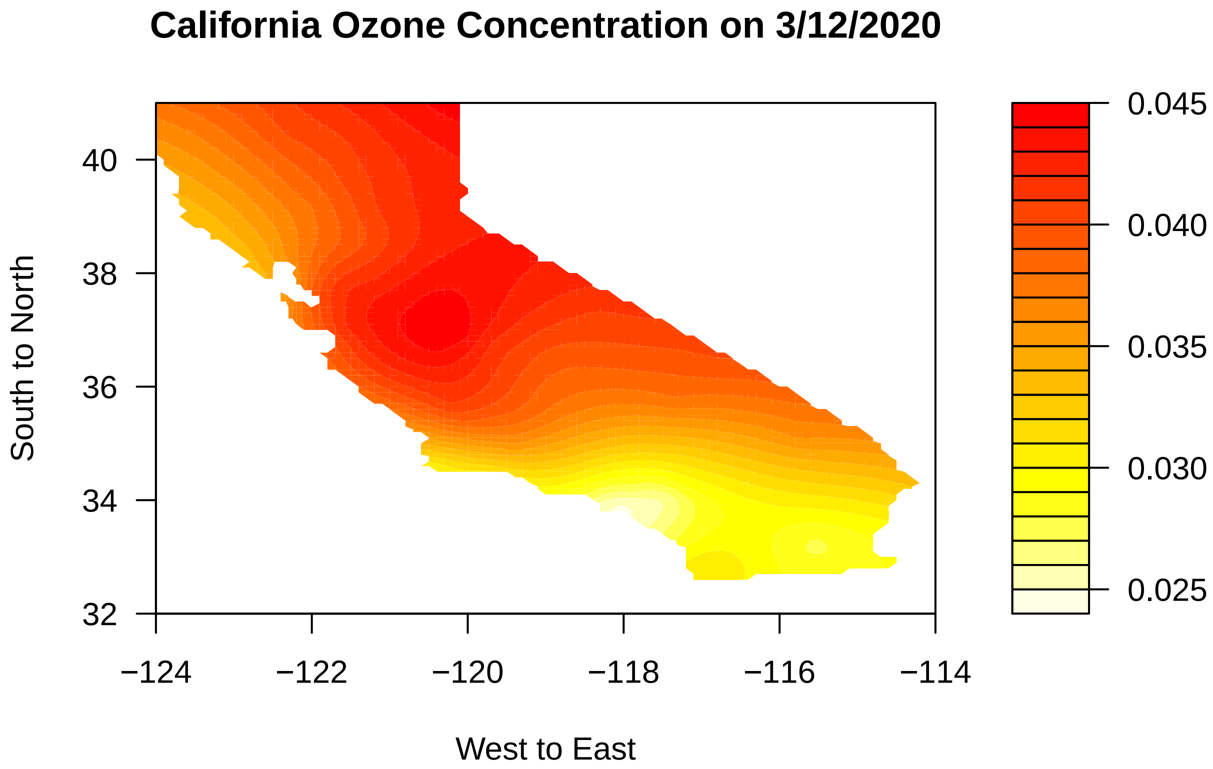
```
#Construct a raster map using the variances of the kriged values:
levelplot(ukriging$data.var~x+y, ukriging, aspect ="iso", main="Universal kriging variances")
```



```
pred <- ukriging$data.pred
qqq <- matrix(pred, length(x), length(y))

#Identify the location of each point in the grid:
in.what.state <- map.where(database="state", x=grd$x, y=grd$y)

#Find the points of the grid that belong to California:
in.us <- which(in.what.state == "california")
#Assign NA values to all the points outside California:
pred[-in.us] <- NA

#The vector pred contains the California points plus the NA values outside California. This vector will
qqq <- matrix(pred, length(x), length(y))

#We can now construct the new raster map:
image(x, y,qqq,
xlab="West to East",ylab="South to North", main="California Ozone Concentration on 3/12/2020")

#And we can add contours:
contour(x, y, qqq, add=TRUE, col="black", labcex=1)
```

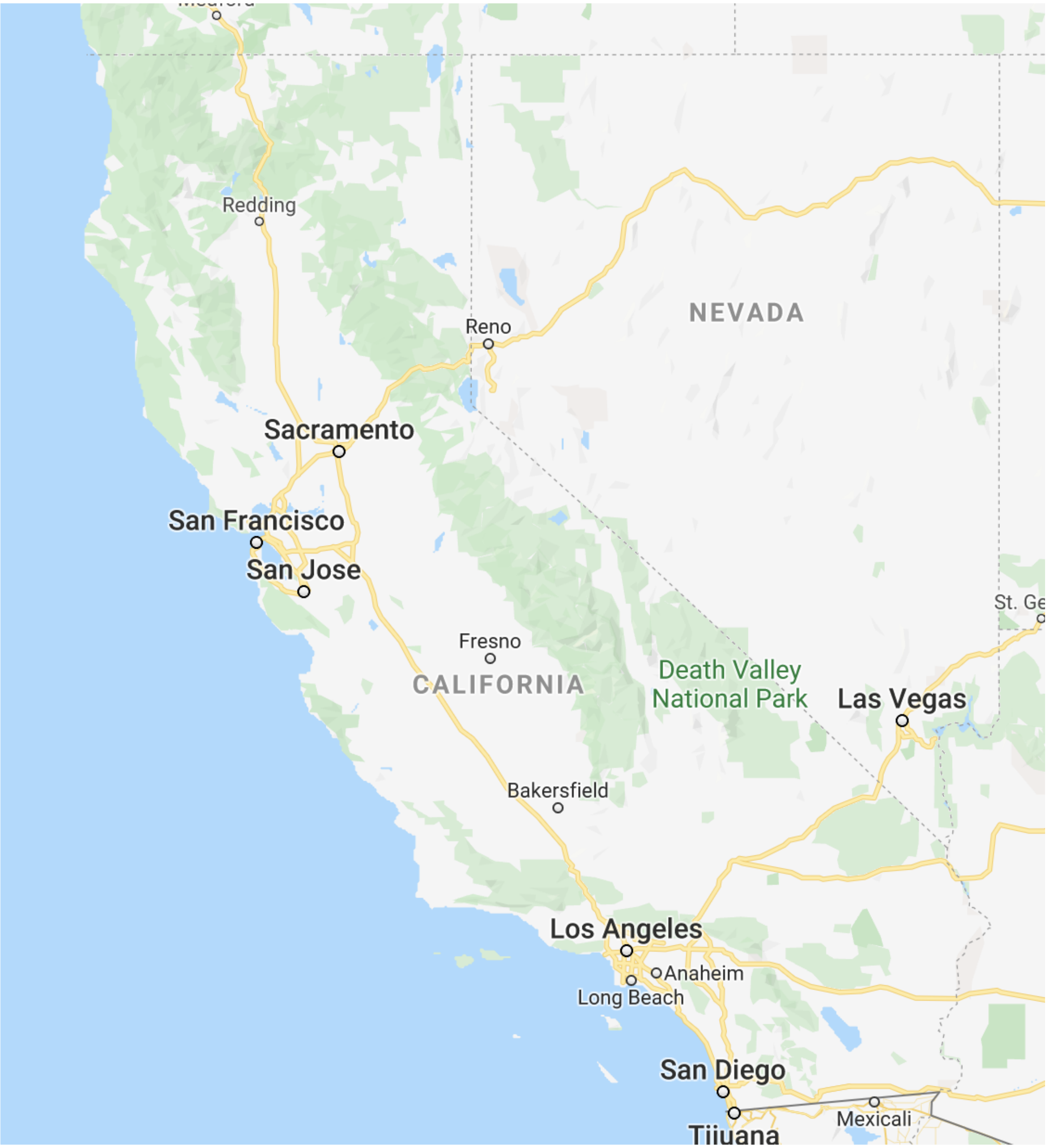


```
#Another function for contours is the following:
filled.contour(x, y,qqq,
xlab="West to East",ylab="South to North", main="California Ozone Concentration on 3/12/2020", col=rev(
```

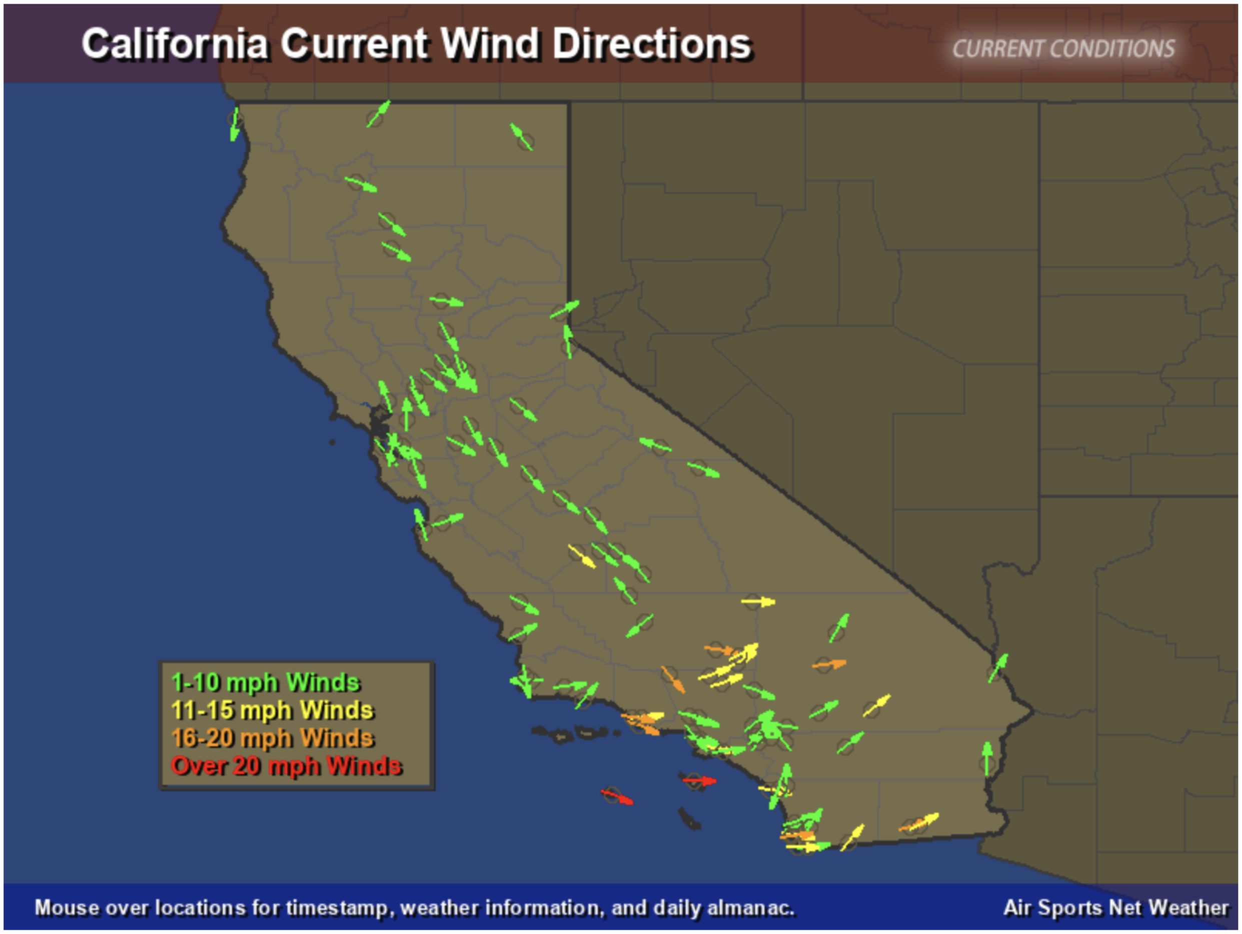


Result

- Since the weather stations are mostly located on the coast, other region's stations are evenly distributed.
- Due to this, prediction variance are lower than 0.00004 throughout the whole region which I am looking for.
- This makes the prediction more reliable.
- Ozone concentraion is heavily clustered between LA and Sanfrancisco.



source from: https://www.google.com/maps/@38.1821923,-117.9869501,6.07z


California Current Wind Directions
CURRENT CONDITIONS
1-10 mph Winds
11-15 mph Winds
16-20 mph Winds
Over 20 mph Winds
Mouse over locations for timestamp, weather information, and daily almanac.
Air Sports Net Weather

source from: http://www.usairnet.com/weather/maps/current/california/wind-direction/

Conclusion

I thought that the ground ozone concentraion will be higher in metropolis areas which are LA, San diego, and San Francisco. However, those areas were not significantly higher than other regions. Highest concentration is observed between the LA and San Francisco. Ozone concentraion on coastal areas are lower than inland areas, and southern california has lower concentration rate than northern california. I think there are more factors that effect ozone concentration. The Ocean can purify the air by absorbing the ozone among the ground atmosphere. Fresh air from the ocean may cause drop of ozone rate on coast area. Heavily concentrated area are surrounded by mountains. Also as we can see in the windmap, Alamo Mountain and MT San Antonio block the wind from proceeding it's way on southern areas. This can cause a giant circulation inside the middle of California. This may keep ozones from spreading from north to south. In further studies, I'll investigate how other pollutants are correlated with the ozone, and see if there's any clustering in those factors. I hope that will explain more on why the Ozone concentration is so much larger in middle of California than the metropolis areas.