Kristina White
Professor Castruccio
ACMS 60855
23 April 2020

Midterm 2: Wolfcamp Aquifer Data

The Wolfcamp Aquifer Dataset contains piezometric head measurements (given in meters) from the Wolfcamp Aquifer in Texas. The data was taken to assess a proposed nuclear waste storage site in Texas. In 1982, under the Nuclear Waste Policy Act, the Department of Energy set out to survey three sites in Nevada, Washington state, and Texas as possible storage sites for highly radioactive nuclear waste. The selected site would eventually contain over 68,000 canisters of radioactive waste underground spaced approximately 30 ft apart in holes or trenches surrounded by salt (Cressie 212). The site must be able to store and isolate the waste for 10,00 years. It takes hundreds of thousands of years for radioactive nuclear waste to decay, and high-level radiation is toxic to humans. Exposure can alter DNA and cause cancer, central nervous system damage, and even be fatal.

For this reason, the possibility of leaks in the cannisters containing the radioactive waste is especially concerning. Radioactive heat could then cause the small amount of water in the salt toward the heat until each cannister is surrounded by water where a chemical reaction between the water and the salt could produce hydrochloric acid (Cressie). This acid could corrode the cannisters. The contaminated water would then move from areas of high pressure to those of low pressure. Given the detrimental health effects of high-level radiation, it is imperative to assess whether the contaminated water from a potential leak could in turn contaminate the water supply of nearby population centers. In the case of the potential Texas site, this would be Amarillo, Texas.

The Wolfcamp Aquifer data contains piezometric head measurements from 85 sites obtained by drilling a narrow pipe through the aquifer as well as the coordinates (in kilometers) of the locations of these sites. A piezometer measures pressure by measuring the height to which water (or other liquid) rises in a column against gravity. For the Wolfcamp Aquifer, each of the 85 pipes were drilled into the aquifer, and the water rose and found its own level in the pipes (Cressie). The piezometric-head measurements (given in meters above sea level) were taken from drill stem tests (Cressie). It is possible the instruments have improved over the years, but given as piezometers are considered reliable equipment, it is likely the data measurements obtained by these instruments is still of high quality.

The bubble plot in Figure 1 shows a trend in the piezometric-head measurements. Pressure seems to decrease as latitude and longitude increase. This is more visible in Figure 2, which shows the piezometric head measurements by longitude and latitude, respectively.

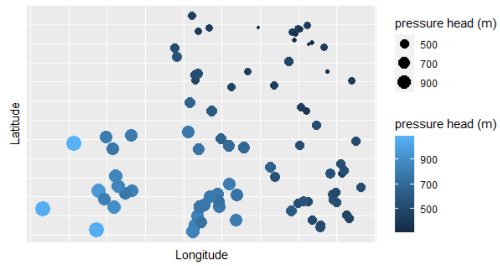
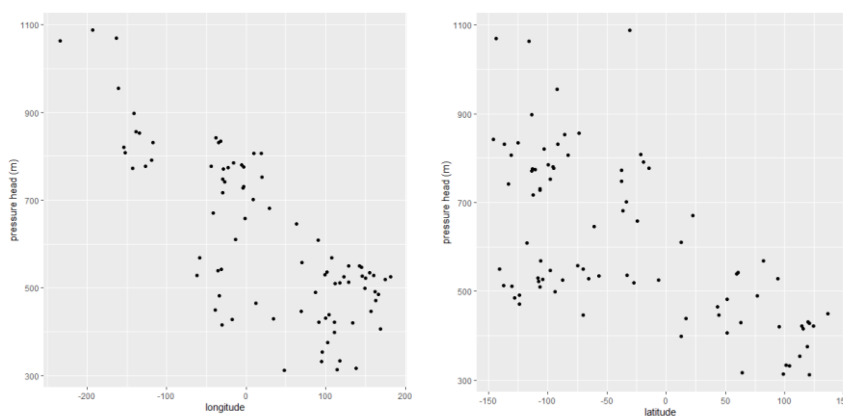


Figure 1 Bubble Plot of piezometric-head measurements



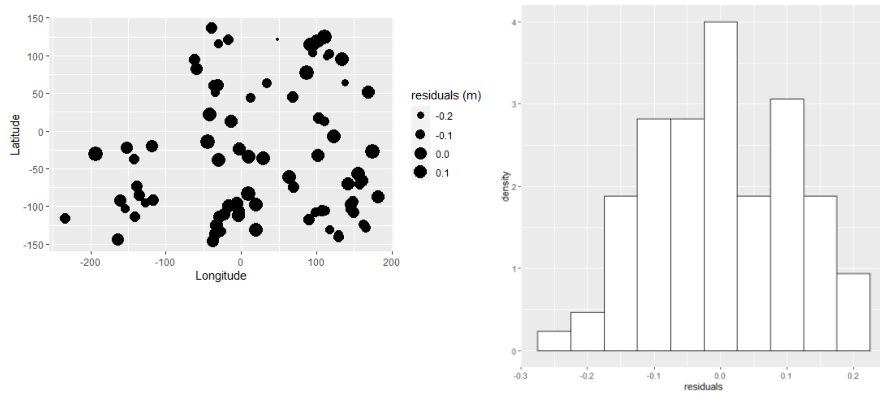
a) longitude vs. piezometric-head measurement *b) latitude vs. piezometric-head measurement*

Figure 2 piezometric-head measurements by longitude and latitude

I fit this trend using a linear model. The model is as follows:

$$\log(z) = \beta_1 x + \beta_2 y + \beta_0 + \varepsilon_s$$

where z is the piezometric head measurement, x is the longitudinal coordinate, y is latitudinal coordinate, β_0 is the intercept, $\varepsilon_s \sim N(0, \mathbf{V})$ is the error term. I chose to use log(z) instead of z because this metric should be positive. Both x and y are significant predictors of log(z) at the $\alpha=0.05$ level (p-value <2e-16). Both x and y have negative coefficients (-0.0019575 and -0.0020547, respectively), indicating a decrease in pressure as longitude and latitude increase. The model has an R^2 value of 0.8815, indicating reasonably good fit. A table of the model parameters is included in Appendix A. The population center of concern for this site is Amarillo, which is in the northeast part of the map where the pressure tends to be lower.



a) Bubble Plot of Residuals   *b) Histogram of Residuals*

Figure 3 Residual Plots

Looking at the residuals of the model, the bubble plot in Figure 3 (a) does not indicate there is spatial dependence among the residuals. The histogram in Figure 3 (b) does indicate there is Gaussianity among the residuals. This means we can assume the residuals are normally distributed.

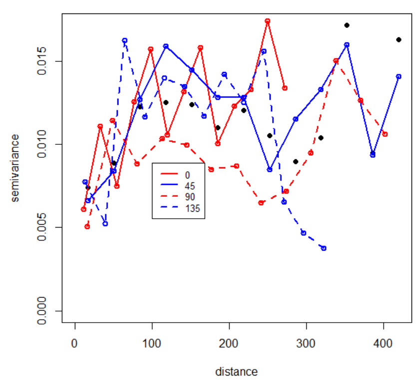
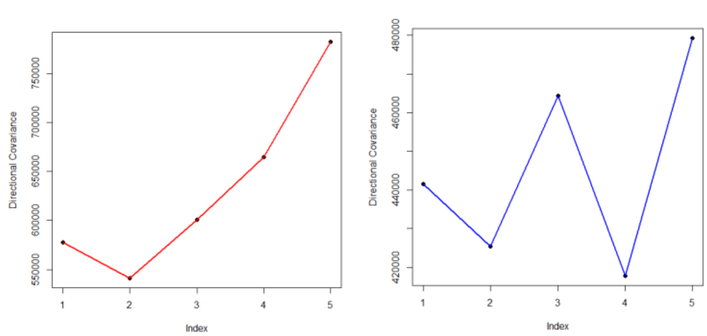


Figure 4 Directional Variogram



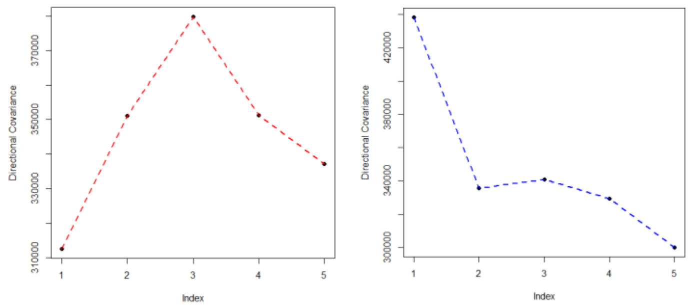
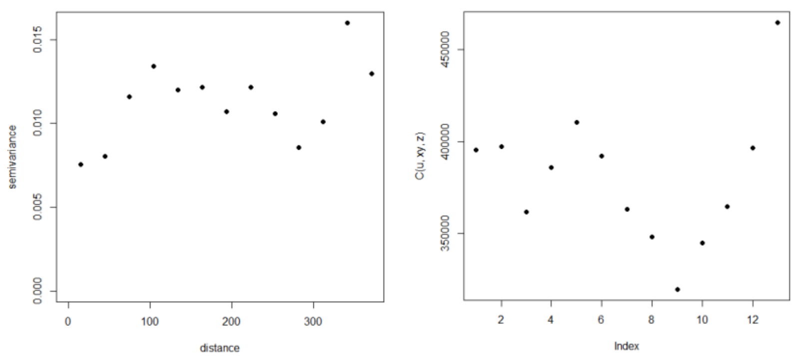


Figure 4 Directional Covariance at 0, 45, 90, and 135 degrees (clockwise from upper left)

Figures 4 and 5 show the directional variogram and covariance at angles of 0, 45, 90, and 135 degrees. The wiggliness in the graphs is in part due to bins where there are fewer points. It does appear that there is a bit of directional dependence at 90 degrees (as the variogram and covariance are at lower values), and at 0 degrees (higher values of covariance). This likely reflects the structure of the aquifer.



a) Empirical Variogram *b) Empirical Covariance*

Figure 6 Empirical Variogram and Covariance. Max distance of 390 used for variogram, while the bins for the covariance are those used for the variogram with 0 included.

Figure 6 included the empirical variogram and the empirical covariance. For the empirical variogram u does not get that large, and the semivariance is also small, meaning Z_{s+u} is close to Z_s. 390 was chosen as the max distance because it is less the maximum distance between any two of the 85 data points, and because it provides enough bins to express a patten without too much wiggliness. There is high covariance among the data points that at first decreases as u increases, then starts to increase with u. Figure 7 shows the fits of the Matern, spherical, and linear variograms. The Matern appears to have the best fit and fits the data best by the origin. I will use the Matern variogram to perform kriging.

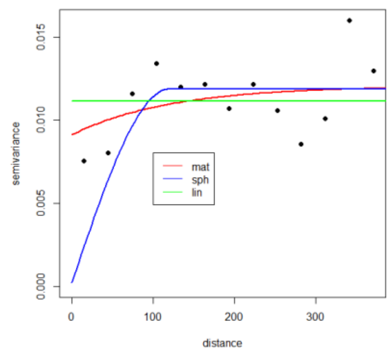


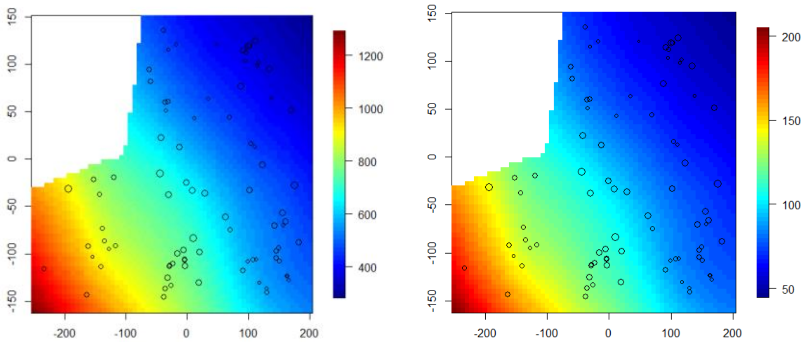
Figure 7 Plot of Matern, Shperical, and Linear Variograms

Kriging allows us to interpolate values for pressure over the grid of points that comprise the aquifer. We first perform kriging on the residuals and then add the linear trend modeled above back in to obtain these estimates on the original scale. We can see the point estimates for the pressure head in Figure 8 (a). They decrease as longitude and latitude increase, which is consistent with the trend we found. The pressure does seem to be lower in the direction of Amarillo, which is located in the northeastern part of the map. This makes sense as water moves from high to low pressure, and we would expect water in an aquifer to move in the direction of a population center to supply water. The uncertainty is higher is places of higher pressure, and we can see the 95% confidence interval for the point estimates in Figure 8 (c) and (d).

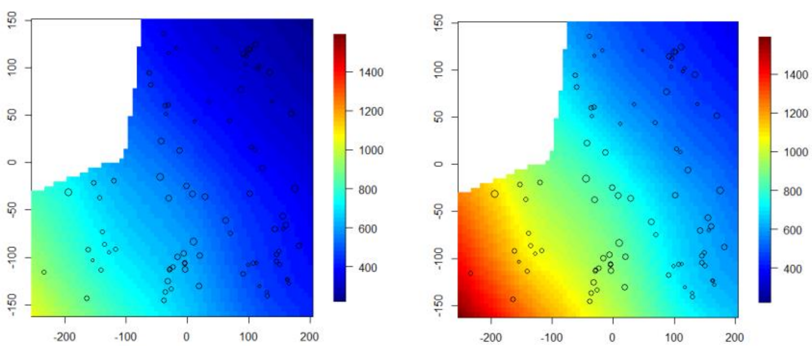
The results are likely reliable. The survey was carried out by scientists at the Department of Energy using high quality instruments. Additionally, given the severe health effects of radiation, the possibility of leaks contaminating a nearby water supply is an issue that would be taken very seriously. Otherwise, the government could face a public health crisis and political and legal consequences. Therefore, the Department of Energy would ensure their data were high quality.

The proposed site in Deaf Smith County (site 15), for example, could pose a problem for Amarillo's water supply (site 59). The difference between their pressure estimates in the log scale is 0.1880282, which is positive and suggests water is flowing from an area of higher pressure in Deaf Smith County to an area of lower pressure in Amarillo. The confidence interval is (0.9934333, 1.4661570). This interval does not include 0, which suggests there is a pressure gradient, and that a potential leak in Deaf Smith County could cause problems for Amarillo's water supply. Based on this conclusion and Amarillo's location in an area with lower water pressure, I would not recommend this site for nuclear waste storage. In 1987, Congress reached the same decision and instead chose the Yucca Mountain site in Nevada.

Finally, this analysis looked at the Wolfcamp Aquifer Data, which contains the coordinates (km) and water pressure measurements (m) for 85 sites at the Wolfcamp Aquifer in Texas. In 1982, the Department of Energy set out to collect this data as part of a survey of three possible nuclear waster storage sites in Nevada, Washington state, and Texas. The chosen site would need to be able to isolate highly radioactive nuclear waste for 10,000 years. This waste remains toxic to humans for hundreds of thousands of years due to the long decay periods of these materials, and the radiation can cause cancer and central nervous system damage, among other health issues. The waste would be stored underground in cannisters surrounded by salt.



a) Kriging Point Estimates *b) Kriging Standard Deviation*



a) Lower Bound *b) Upper Bound*

Figure 8 Kriging Results in the Original Scale

However, potential leaks in these cannisters could cause water and salt to react to corrode the cannisters, and contaminated water could then move from areas of high pressure to areas of low pressure. It is important to see if these sites could lead to the contamination of a population center's water supply in the event of a leak. The pressure measurements were collected using a piezometer, which measures the height a column of water rises to against gravity. When analyzing this data, the location was a good predictor of the pressure measurement, and Amarillo is located in an area of lower water pressure. When looking specifically at the Deaf Smith county site, there was a pressure difference so contaminated water could flow from this site to Amarillo. In 1987, Congress rejected the proposed Texas site and chose the Yucca Mountain site in Nevada as a nuclear waste storage facility.

Appendix A

Table 1.1 Linear Model Estimates	
Parameter	Estimate
β_0	6.3536945
β_1	-0.0019575
β_2	-0.0020547

Estimates highlighted in yellow are significant at the α=0.05 level.

Appendix B

These are plots for the kriged values and kriging standard deviations of the residuals. The point estimates plot is fairly smooth, but there is not a practical interpretation for the estimates of the residuals.

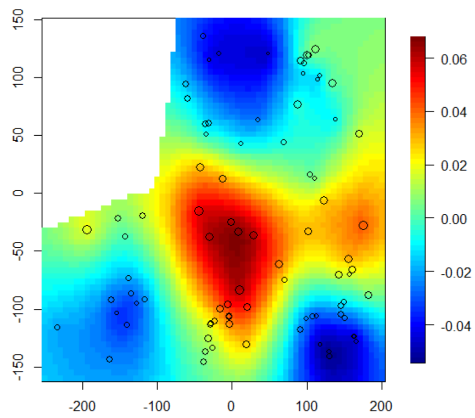


Figure 9 Kriging Point Estimates for the Residuals

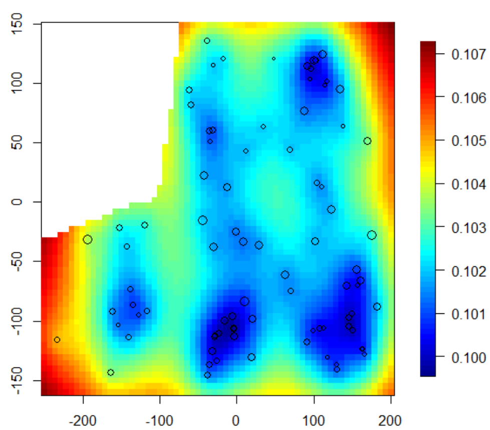


Figure 10 Kriging Standard Deviations for the Residuals

Appendix C

R code

```
library(geostatsp)
library(geoR)
library(fields)
library(tidyverse)
library(sp)

data("wolfcamp")

load("borders.Rdata")

geodata <- wolfcamp

z <- geodata$data
x <- geodata$coords[, 1]
y <- geodata$coords[, 2]


df=data.frame(lat=y,long=x, Z=z)
p=ggplot()+geom_point( data=df, aes(x=x, y=y, size = Z, color=Z))
p=p + labs(x = "Longitude",y = "Latitude",size = "pressure head (m)",colour = "pressure head (m)")+ coord_fixed()
p=p + theme(axis.text.x = element_blank(),axis.text.y = element_blank(),axis.ticks = element_blank())
p
```

```
df=data.frame(X=z)

df2=data.frame(X=log(z))

ggplot()+geom_histogram(data=df, aes(x=X,y=..density..), binwidth=75,color="black", fill="white")+xlab("pressure head (m)")

ggplot()+geom_histogram(data=df2, aes(x=X,y=..density..),binwidth=0.25,color="black", fill="white")+xlab("log pressure head (log m)")



df=data.frame(X=x,Y=y,Z=z)

ggplot(df,aes(X, Z))+geom_point()+ylab("pressure head (m)")+xlab("longitude")

ggplot(df,aes(Y, Z))+geom_point()+ylab("pressure head (m)")+xlab("latitude")



mod.log.xy=lm(log(z)~x+y)

summary(mod.log.xy)



## storing the data

geodata$data.orig=geodata$data

geodata$data=residuals(mod.log.xy)



## setting up the prediction grid

gr = pred_grid(borders,by=3)

gr.bor=locations.inside(loc=gr, borders=borders)
```

```
gr.pred=data.frame(x=gr.bor[,1],y=gr.bor[,2])

pred=predict(mod.log.xy,newdata=gr.pred,interval="prediction",level=0.95)
geodata$trend=pred[,1]
sd_pred=(pred[,3]-pred[,1])/1.96

## plotting residuals

df=data.frame(lat=geodata$coords[, 2],long=geodata$coords[, 1],Z=geodata$data)
p=ggplot()+geom_point( data=df, aes(x=long, y=lat, size = Z))
p=p + labs(x = "Longitude",y = "Latitude",size = "residuals (m)")+ coord_fixed()
p

df=data.frame(X=geodata$data)
ggplot()+geom_histogram(data=df, aes(x=X,y=..density..),color="black", binwidth= 0.05,
fill="white")+xlab("residuals")

max(dist(geodata$coords))
md = 437
lz.v = variog(geodata,max.dist=md)
lz.v.a0 = variog(geodata,max.dist=md,direction=0)
lz.v.a45 = variog(geodata,max.dist=md,direction=pi/4)
lz.v.a90 = variog(geodata,max.dist=md,direction=pi/2)
lz.v.a135 = variog(geodata,max.dist=md,direction=3/4*pi)
```

```
plot(lz.v,pch=19)

lines(lz.v.a0,col="red",lwd=2)

lines(lz.v.a45,col="blue",lwd=2)

lines(lz.v.a90,col="red",lty=2,lwd=2)

lines(lz.v.a135,col="blue",lty=2,lwd=2)

legend(100,lz.v$v[2],col=c("red","blue","red","blue"),lty=c(1,1,2,2),c(0,45,90,135),lwd=c(2,2,2,
2),cex=0.8)



xy = cbind(x,y)


u =  seq(0, 250, 50)


theta = 0


t = 0.393


C <- function(u, xy, z, theta, t){


  n <- length(u)


  c <- rep(0, n-1)


  Nk <- rep(0, n-1)


  for(j in 1:n-1){


    for(i in 1:length(z)){
```

```
      for (m in 1:length(z)){
        d = sqrt(((xy[i, 1]-xy[m, 1])^2)+((xy[i, 2]-xy[m, 2])^2))
        if(i != m){
          a = acos((c(xy[i,1], xy[i,2])%*%c(xy[m,2],xy[m,2]))/(norm(xy[i, ])*norm(xy[m, ])))
        }
        flag = (d > u[j]) && (d < u[j+1]) && (a > (theta-t)) && (a <= (theta+t))

        if(flag){
          c[j] <- c[j] + sum(z[i]*z[m])
          Nk[j] <- Nk[j] + sum(flag)
        }
      }
    }
  }

  return(c/Nk)

}

dcv1 <- C(u, xy, z, theta, t)
dcv2 <- C(u, xy, z, pi/4, t)
dcv3 <- C(u, xy, z, pi/2, t)
dcv4 <- C(u, xy, z, 3*pi/4, t)

plot(dcv1, pch = 19, ylab = "Directional Covariance")
lines(dcv1,col="red",lwd=2)
plot(dcv2, pch = 19, ylab = "Directional Covariance")
```

```
lines(dcv2,col="blue",lwd=2)
plot(dcv3, pch = 19, ylab = "Directional Covariance")
lines(dcv3,col="red",lty=2,lwd=2)
plot(dcv4, pch = 19, ylab = "Directional Covariance")
lines(dcv4,col="blue",lty=2,lwd=2)




geodata.vgm = variog(geodata,max.dist=390)
geodata.vgm$n
plot(geodata.vgm, pch = 19)


u =  c(0, geodata.vgm$u)


C <- function(u, xy, z){
n_u=length(u)
C=rep(0,n_u-1)
Nk=rep(0,n_u-1)
n = length(z)

for (j in 1:(n_u-1)){
  for (i in 1:n){
    d=sqrt((xy[i,1]-xy[i:n,1])^2+(xy[i,2]-xy[i:n,2])^2)
    flag=(d>=u[j])&(d<u[j+1])
    if (sum(flag)>0){
```

```
      C[j]=C[j]+sum(z[i]*z[i:n][flag])
      Nk[j]=Nk[j]+sum(flag)
    }
  }
}
return(C/Nk)
}

C(u, xy, z)
plot(C(u, xy, z), pch=19)



geodata.vgm = variog(geodata,max.dist=390)
geodata.vgm$n
plot(geodata.vgm, pch = 19)

### 2) Fit variogram model, with nugget
geodata.fit.mat =
variofit(geodata.vgm,ini=c(var(geodata$data),120),cov.model="matern",fix.nugget =
F,kap=0.5,fix.kappa = T)
geodata.fit.sph = variofit(geodata.vgm,ini=c(var(geodata$data),120),cov.model="spherical")
geodata.fit.lin = variofit(geodata.vgm,ini=c(var(geodata$data),120),cov.model="linear")



### 3) Plot for visual inspection
plot(geodata.vgm,pch=19)
lines(geodata.fit.mat,col="red",lwd=2)
lines(geodata.fit.sph,col="blue",lwd=2)
lines(geodata.fit.lin,col="green",lwd=2)
```

```
legend(100,geodata.vgm$v[2],lty=1,col=c("red","blue","green"),c("mat","sph","lin"))



## kriging commands
KC = krige.control(type="sk",obj.model = geodata.fit.mat)
sk =  krige.conv(geodata, loc = gr, borders = borders,krige=KC)

############ residual plots



## plotting kriging point estimates
## remember these are the RESIDUALS!
quilt.plot(gr.bor[,1],gr.bor[,2], sk$predict)
points(geodata,pch=,add=T)

## plotting kriging standard deviation
## remember these are the RESIDUALS!
quilt.plot(gr.bor[,1],gr.bor[,2], sqrt(sk$krige.var))
points(geodata,pch=,add=T)




## Confidence intervals
Z_CI_l=exp(geodata$trend+sk$predict-1.96*sqrt(sk$krige.var))
Z_CI_u=exp(geodata$trend+sk$predict+1.96*sqrt(sk$krige.var))
zlim_l=min(Z_CI_l)
zlim_u=max(Z_CI_u)
```

```
## plotting kriging lower and upper bound
quilt.plot(gr.bor[,1],gr.bor[,2], Z_CI_l,zlim=c(zlim_l,zlim_u))
points(geodata,pch=,add=T)

## plotting kriging lower and upper bound
quilt.plot(gr.bor[,1],gr.bor[,2], Z_CI_u,zlim=c(zlim_l,zlim_u))
points(geodata,pch=,add=T)

## point estimates
prect.pred=exp(geodata$trend+sk$predict+sk$krige.var/2)

quilt.plot(gr.bor[,1],gr.bor[,2], prect.pred)
points(geodata,pch=,add=T)


# plotting the variance in the original scale
mu=geodata$trend+sk$predict
sigma=sqrt(sd_pred^2+sk$krige.var)
or.var=(exp(sigma^2)-1)*exp(2*mu+sigma^2)

quilt.plot(gr.bor[,1],gr.bor[,2], sqrt(or.var))
points(geodata,pch=,add=T)

d1 <- c ()
d2 <- c()
for(i in 1:length(gr.bor[,1])){
  d1[i] <- sqrt(((xy[15, 1]-gr.bor[i, 1])^2)+((xy[15, 2]-gr.bor[i, 2])^2))
```

```
  d2[i] <- sqrt(((xy[59, 1]-gr.bor[i, 1])^2)+((xy[59, 2]-gr.bor[i, 2])^2))
}

which.min(d1)
which.min(d2)

est = (geodata$trend[8621]+sk$predict[8621])-(geodata$trend[9327]+sk$predict[9327])
d = sqrt(((gr.bor[8621, 1]-gr.bor[9327, 1])^2)+((gr.bor[8621, 2]-gr.bor[9327, 2])^2))
uvec = seq(0, 405.5534, 57.93962)
geodata.vgm2 = variog(geodata,option = "bin",uvec=uvec)
geodata.fit.mat2 = variofit(geodata.vgm2,ini=c(var(geodata$data),50),cov.model="matern",fix.nugget = F,kap=0.5,fix.kappa = T)
#var = geodata.fit.mat2$cov.pars[1]
var = geodata.vgm2$v[2]
Z_CI_l=exp(est-1.96*sqrt(var))
Z_CI_u=exp(est+1.96*sqrt(var))
c(Z_CI_l, Z_CI_u)
```

References

Cressie, N.A.C (1993) Statistics for Spatial Data. New York: Wiley.