2023 554 R Notes on Mapping for Point Data

Jon Wakefield
Departments of Biostatistics and Statistics
University of Washington

2023-03-06

Overview

In these notes we will consider mapping and modeling of point data in which the (nominal) exact locations are known.

We will look at modeling a spatially-indexed continuous response via:

- Conventional Kriging via MLE and variants
- A generalized additive model (GAM)
- A Bayesian approach using stochastic partial differential equations (SPDE)

Continuous Response: Motivating Example

We illustrate methods for continuous data using on Zinc levels in the Netherlands.

This data set gives locations and top soil heavy metal concentrations (in ppm), along with a number of soil and landscape variables, collected in a flood plain of the river Meuse, near the village Stein in the South of the Netherlands.

Heavy metal concentrations are bulk sampled from an area of approximately 28km $\times$ 39km.

The Meuse data are in a variety of packages. The version in the geoR library are not a spatial object, but can be used with likelihood and Bayes methods.

Meuse analysis using `geostat` functions

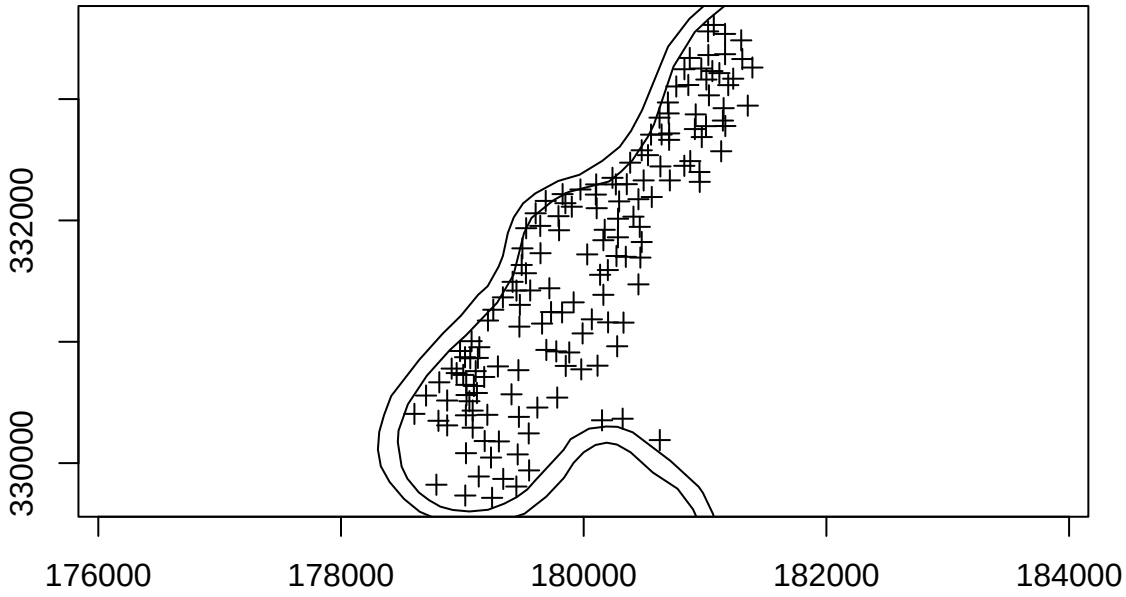
We look at the sampling locations and then examine variograms.

```
library(tidyverse)
library(ggpubr)
library(viridis)
library(geoR)
data("meuse")
library(sp)
pal <-  function(n = 9){ brewer.pal(n, "Reds") }
data(meuse)
```

```
coords <- SpatialPoints(meuse[,c("x","y")])
meuse1 <- SpatialPointsDataFrame(coords,meuse)
data(meuse.riv)
river_polygon <- Polygons(list(Polygon(meuse.riv)),ID="meuse")
rivers <- SpatialPolygons(list(river_polygon))
coordinates(meuse) = ~x+y
```

Zinc: Sampling locations

```
plot(meuse1,axes=T)
plot(rivers,add=T)
```



Exploratory analysis

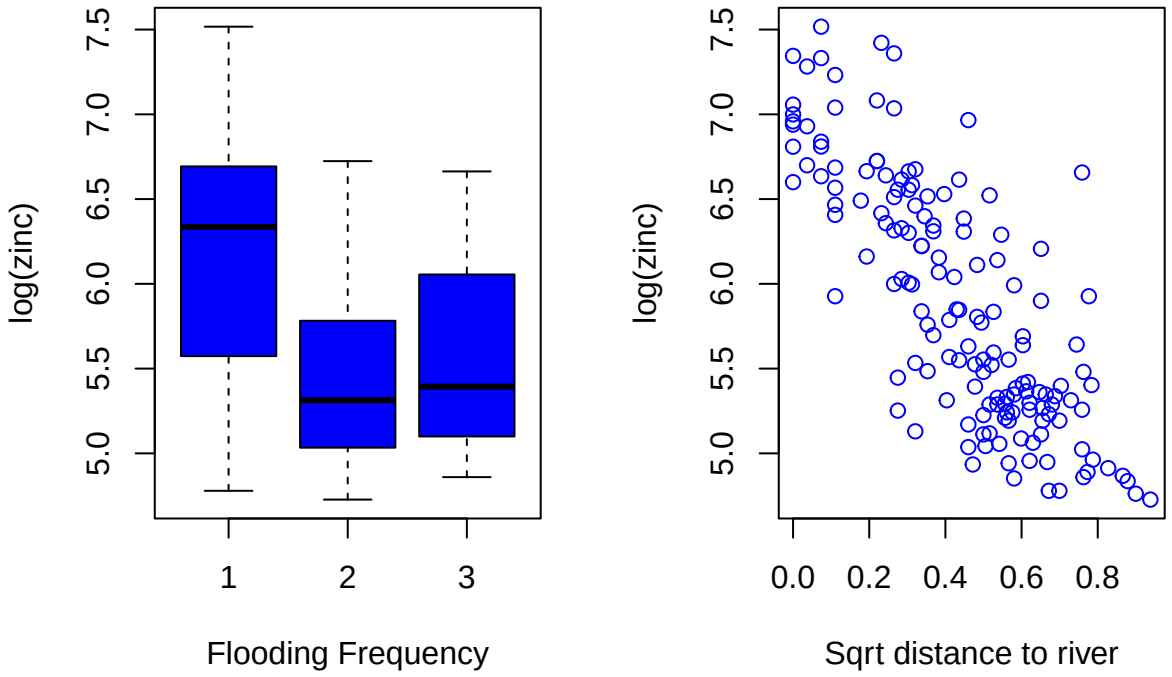
We work with log(zinc) as the distribution is more symmetric than on the original scale, and the variance more constant across levels of covariates.

It's often a good idea to do some exploratory data analysis (EDA), so let's see how log(zinc) varies by two possible covariates:

- Flooding frequency (`ffreq`); 1 = once in two years; 2 = once in ten years; 3 = one in 50 years
- Distance to the Meuse river (`dist`); normalized to $[0, 1]$

We focus on these, since they are available across the study region and so can be used for prediction. Following previous authors, we take `sqrt(dist)` which is closer to linearity.

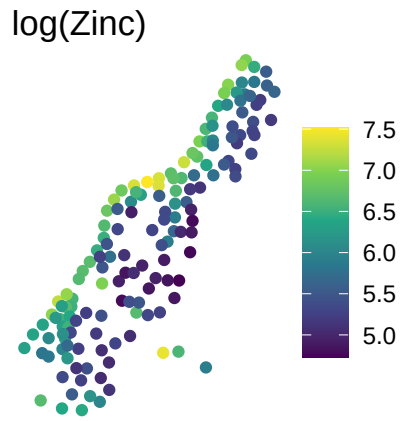
```
par(mfrow=c(1,2))
plot(log(meuse$zinc)~meuse$ffreq,ylab="log(zinc)",xlab="Flooding Frequency",col="blue")
plot(log(meuse$zinc)~sqrt(meuse$dist),ylab="log(zinc)",xlab="Sqrt distance to river",col="blue")
```

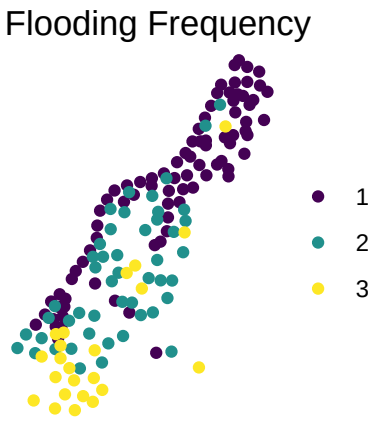


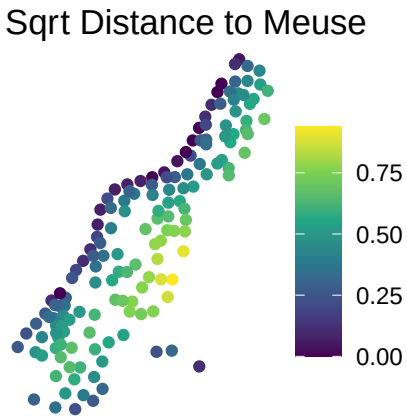

```
# plot(log(meuse$zinc)~meuse$elev,ylab="log(zinc)",xlab="Elevation",col="blue")
```

Also map log(zinc) and these covariates.

```
m.sf <- sf::st_as_sf(meuse, coords = c("x","y"))
m.sf$logzinc <- log(m.sf$zinc)
m.sf$sdist <- sqrt(m.sf$dist)

a <- ggplot() + geom_sf(data = m.sf[,"logzinc"], aes(color = logzinc)) +
  theme_void() + scale_color_viridis_c() + labs(title = "log(Zinc)", color=NULL)
b <- ggplot() + geom_sf(data = m.sf[,"ffreq"], aes(color = ffreq)) +
  theme_void() + scale_color_viridis(discrete=T) + labs(title = "Flooding Frequency", color=NULL)
c <- ggplot() + geom_sf(data = m.sf[,"sdist"], aes(color = sdist)) +
  theme_void() + scale_color_viridis_c() + labs(title = "Sqrt Distance to Meuse", color=NULL)
# d <- ggplot() + geom_sf(data = m.sf[,"elev"], aes(color = elev)) + theme_void() + scale_color_viridis
ggpubr::ggarrange(a,b,c, nrow=2, ncol=2)
```

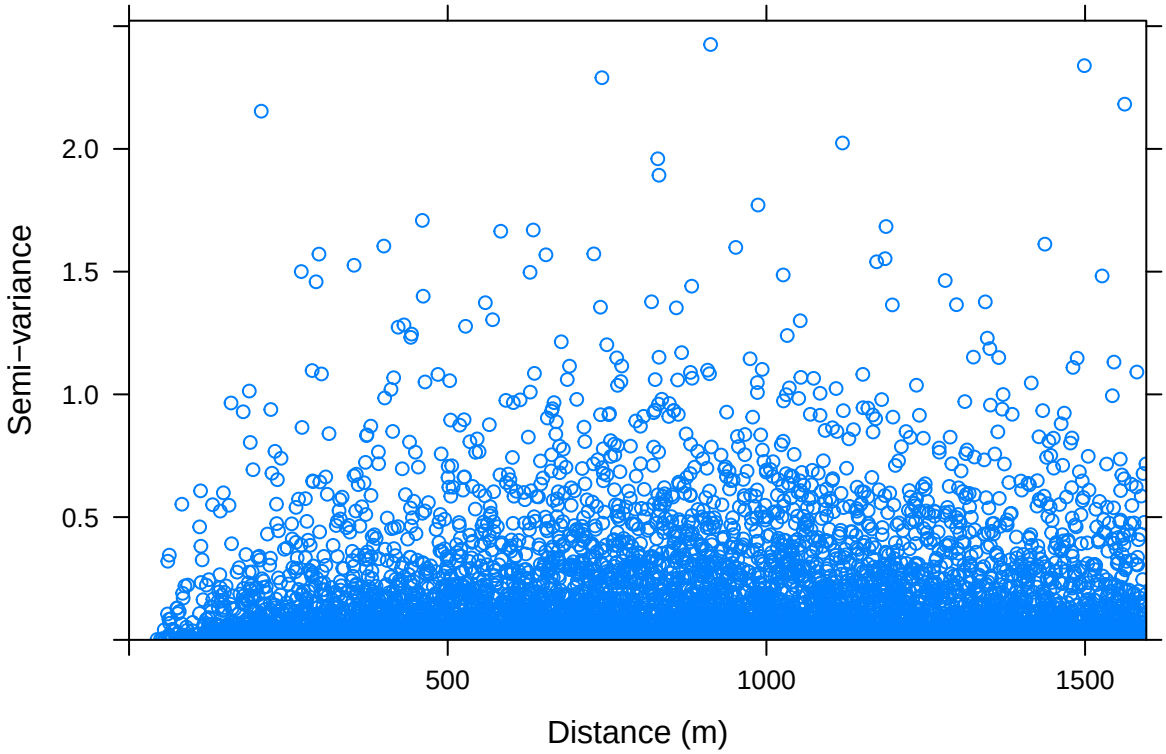






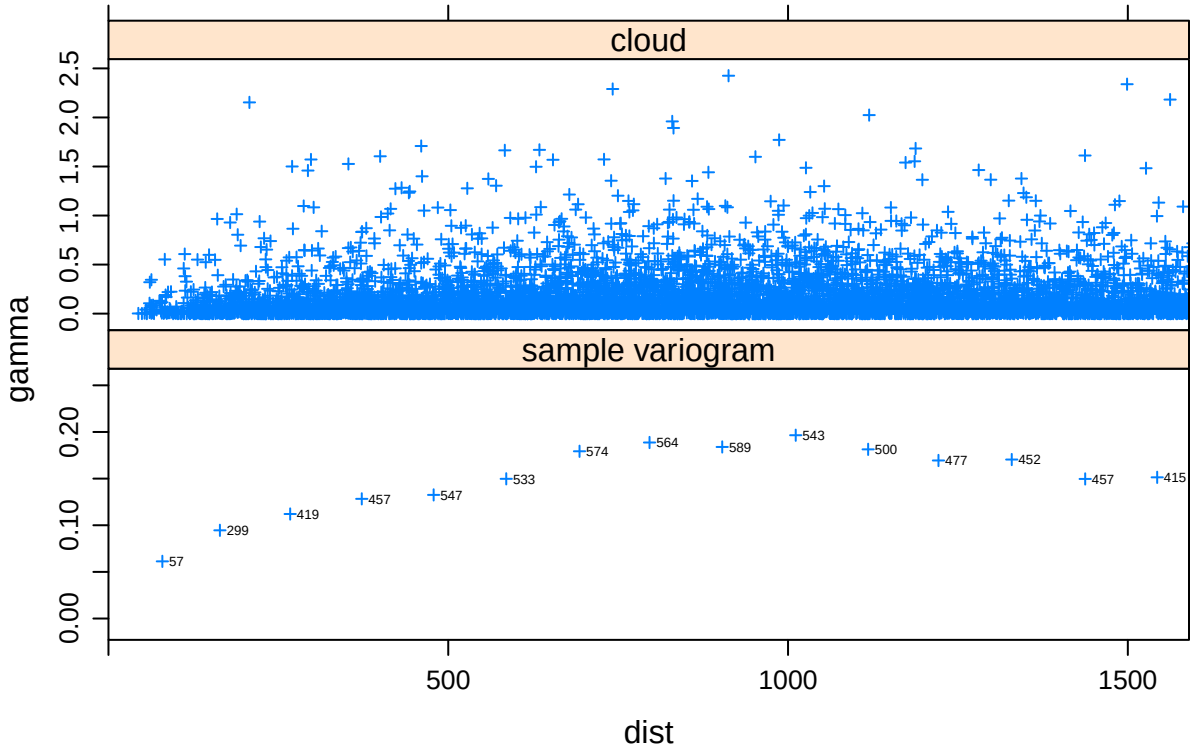
log(zinc): Variogram cloud, tremd removed

```
library(gstat)
cld <- variogram(log(zinc) ~ sqrt(meuse$dist)+as.factor(meuse$ffreq), meuse, cloud = TRUE)
plot(cld,ylab="Semi-variance",xlab="Distance (m)")
```



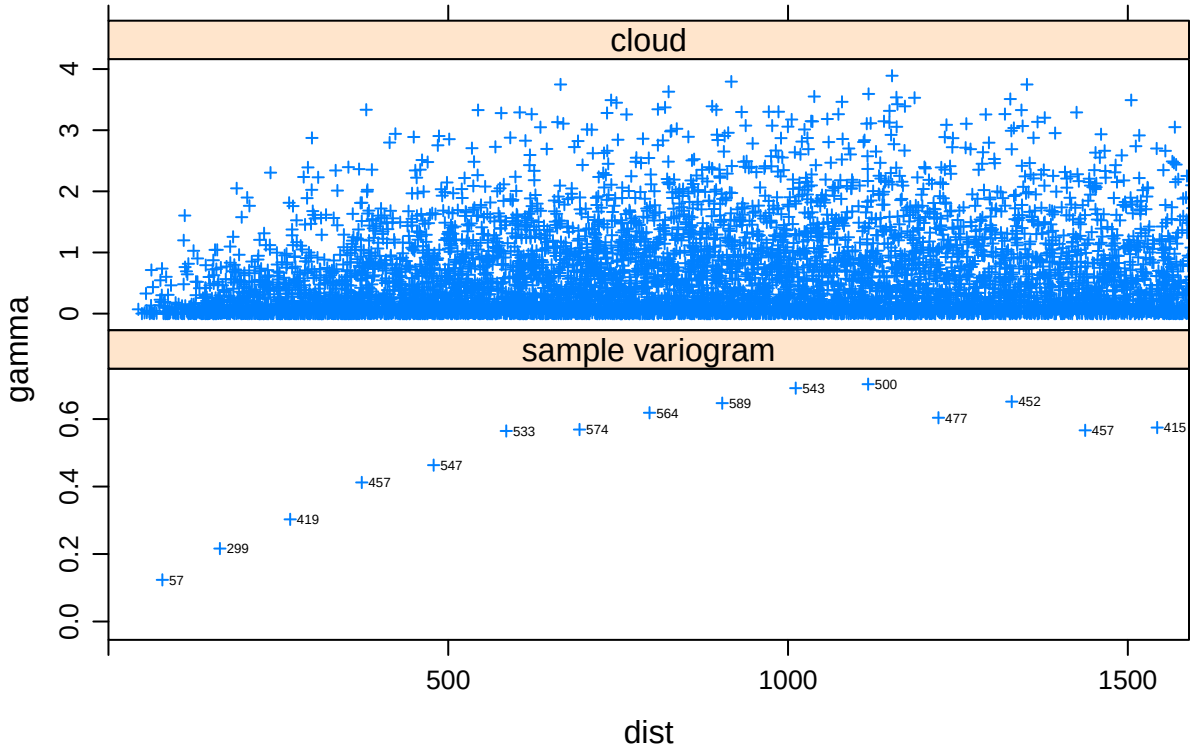
More variograms, with sample sizes

```
library(lattice)
cld <- variogram(log(zinc) ~ sqrt(meuse$dist)+as.factor(meuse$ffreq), meuse, cloud = TRUE)
svgm <- variogram(log(zinc) ~ sqrt(meuse$dist)+as.factor(meuse$ffreq), meuse)
d <- data.frame(gamma = c(cld$gamma, svgm$gamma),
    dist = c(cld$dist, svgm$dist),
    id = c(rep("cloud", nrow(cld)), rep("sample variogram", nrow(svgm)))
    )
xyplot(gamma ~ dist | id, d,
    scales = list(y = list(relation = "free",
      #ylim = list(NULL, c(-.005,0.25))),
      limits = list(NULL, c(-.005,0.25)))),
    layout = c(1, 2), as.table = TRUE,
    panel = function(x,y, ...) {
        if (panel.number() == 2)
            ltext(x+10, y, svgm$np, adj = c(0,0.5),cex=.4) #$
        panel.xyplot(x,y,...)
    },
    xlim = c(0, 1590),
    cex = .5, pch = 3
)
```



More variograms

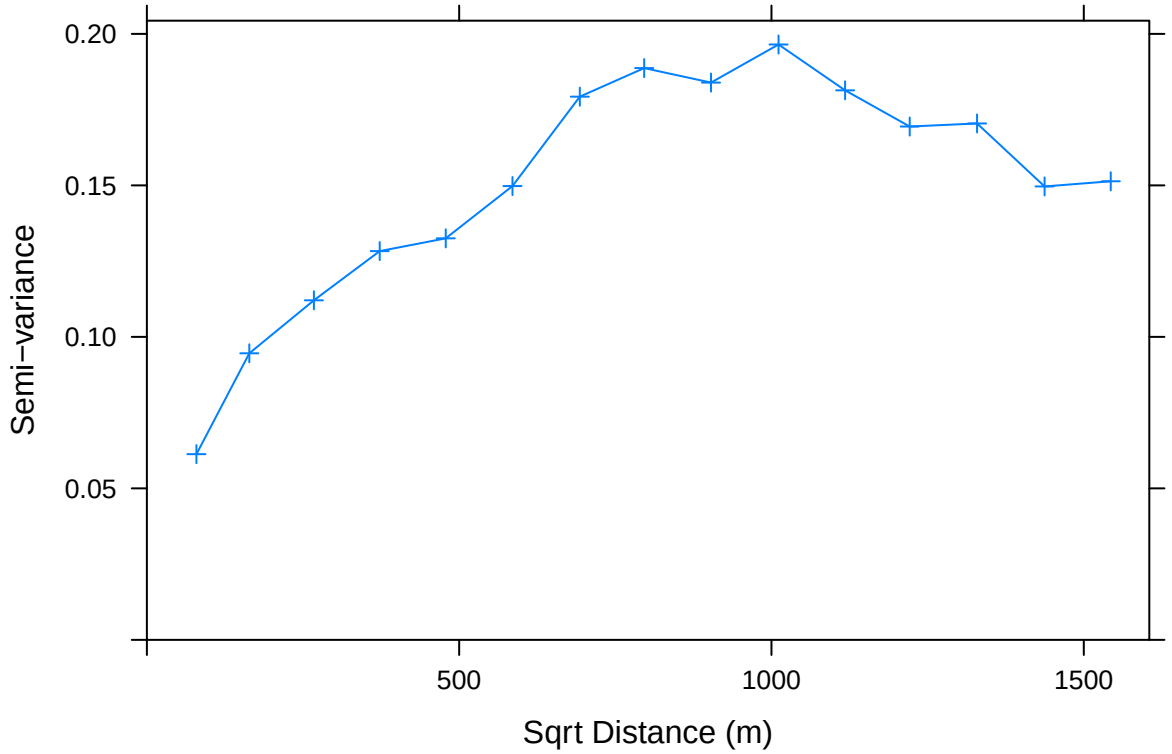
```
cld <- variogram(log(zinc) ~ 1, meuse, cloud = TRUE)
svgm <- variogram(log(zinc) ~ 1, meuse)
d <- data.frame(gamma = c(cld$gamma, svgm$gamma),
    dist = c(cld$dist, svgm$dist),
    id = c(rep("cloud", nrow(cld)), rep("sample variogram", nrow(svgm)))
    )
xyplot(gamma ~ dist | id, d,
    scales = list(y = list(relation = "free",
      #ylim = list(NULL, c(-.005,0.7))),
      limits = list(NULL, c(-.005,0.7)))),
    layout = c(1, 2), as.table = TRUE,
    panel = function(x,y, ...) {
        if (panel.number() == 2)
            ltext(x+10, y, svgm$np, adj = c(0,0.5),cex=.4) #$
        panel.xyplot(x,y,...)
    },
    xlim = c(0, 1590),
    cex = .5, pch = 3
)
```



Monte Carlo simulations of semi-variogram

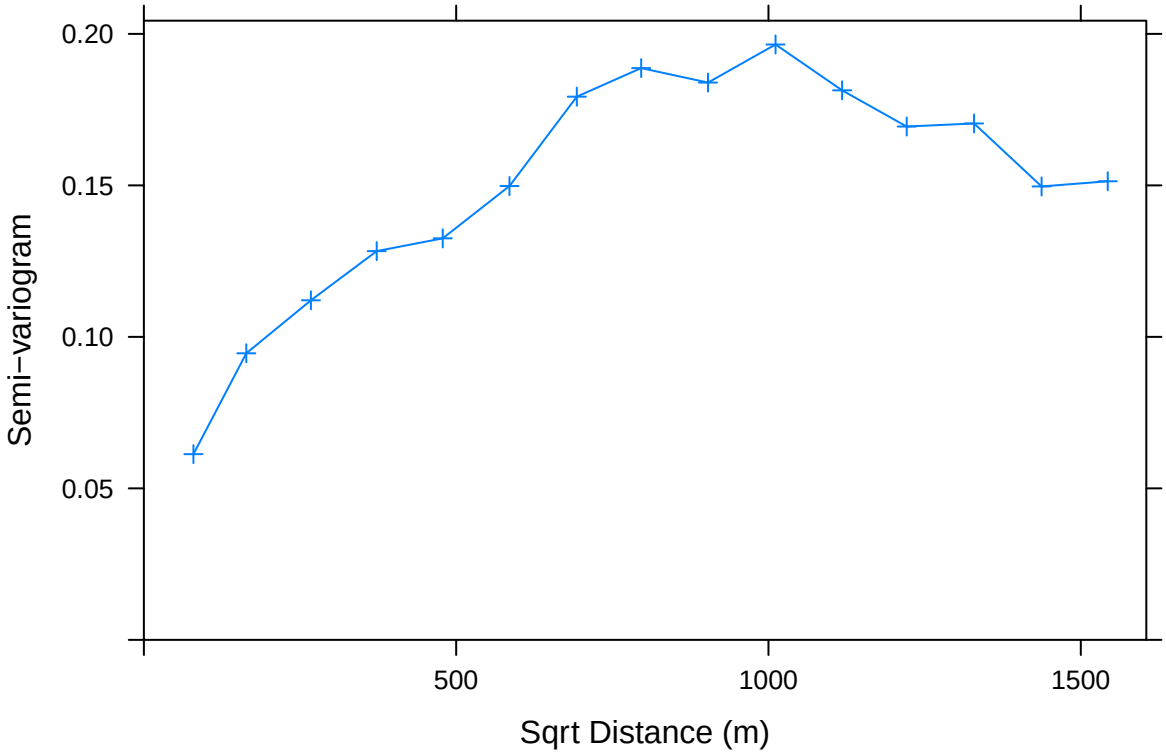
We simulate 100 datasets with random relabeling of points, and then form variograms for each.

```
v <- variogram(log(zinc) ~ sqrt(meuse$dist)+as.factor(meuse$ffreq), meuse)
plot(v, type = 'b', pch = 3,xlab="Sqrt Distance (m)",ylab="Semi-variance")
fn = function(n = 100) {
        for (i in 1:n) {
           meuse$random = sample(meuse$zinc)
           v = variogram(log(random) ~ 1, meuse)
           trellis.focus("panel", 1, 1, highlight = FALSE)
           llines(v$dist, v$gamma, col = 'grey')
           trellis.unfocus()
        }
}
fn()
```



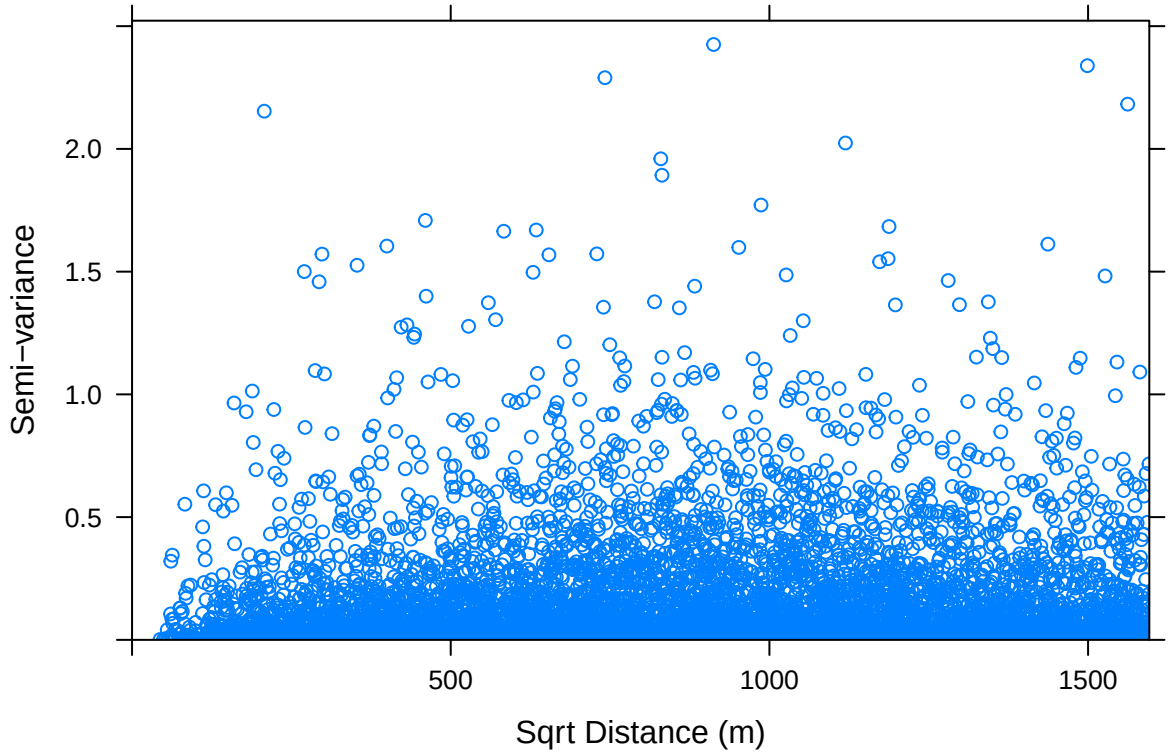
Monte Carlo envelopes under no spatial dependence - it is clear there is dependence here.

```
v <- variogram(log(zinc) ~ sqrt(meuse$dist)+as.factor(meuse$ffreq), meuse)
plot(v, type = 'b', pch = 3,xlab="Sqrt Distance (m)",ylab="Semi-variogram")
fn = function(n = 100) {
        for (i in 1:n) {
           meuse$random = sample(meuse$zinc)
           v = variogram(log(random) ~ 1, meuse)
           trellis.focus("panel", 1, 1, highlight = FALSE)
           llines(v$dist, v$gamma, col = 'grey')
           trellis.unfocus()
        }
}
fn()
```



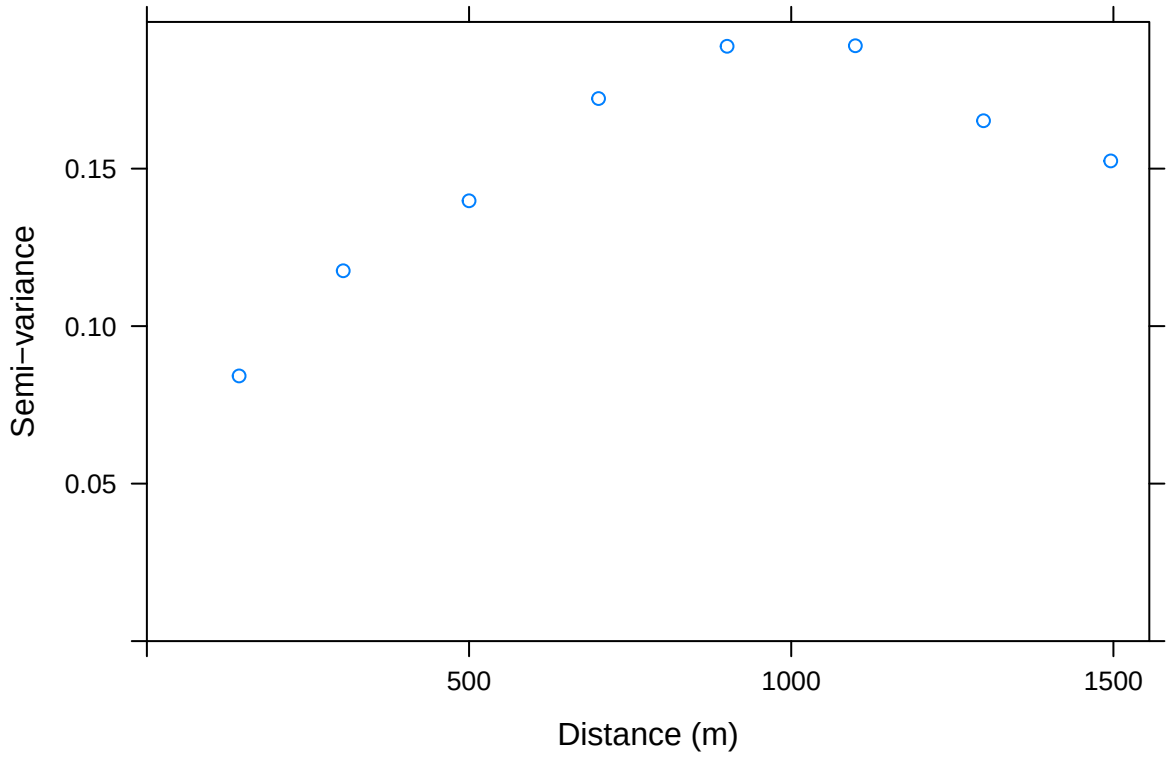
log(zinc): Variogram cloud, detrended

```
cld2 <- variogram(log(zinc) ~ sqrt(meuse$dist)+as.factor(ffreq), meuse, cloud = TRUE)
plot(cld2,ylab="Semi-variance",xlab="Sqrt Distance (m)")
```



Binned variogram for log(zinc), detrended

```
gstatbin <- variogram(log(zinc) ~ sqrt(meuse$dist)+as.factor(ffreq), meuse, width=200)
plot(gstatbin,ylab="Semi-variance",xlab="Distance (m)")
```



Directional variogram with linear trend removed

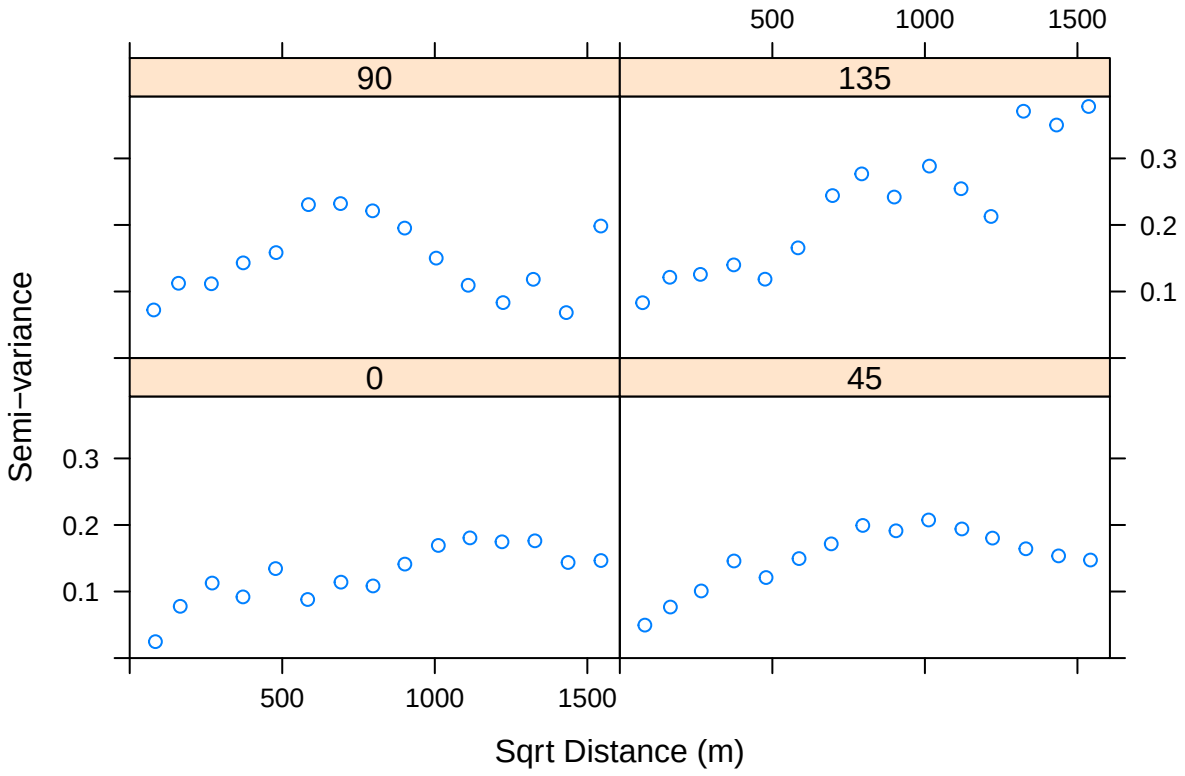
We form 4 variograms with data taken from different directions, with 0 and 90 corresponding to north and east, respectively.

Note that 0 is the same as 180.

```
dircld <- variogram(log(zinc)~sqrt(meuse$dist)+as.factor(ffreq), meuse, alpha=c(0,45,90,135))
```

Directional variogram with linear trend removed

```
plot(dircld,xlab="Sqrt Distance (m)",ylab="Semi-variance")
```



Other capabilities in `gstat`

See

- `fit.variogram` for estimation from the variogram
- `krige` (and associated functions) for Kriging,
- `vgm` generates variogram models

`geoR` for geostatistics

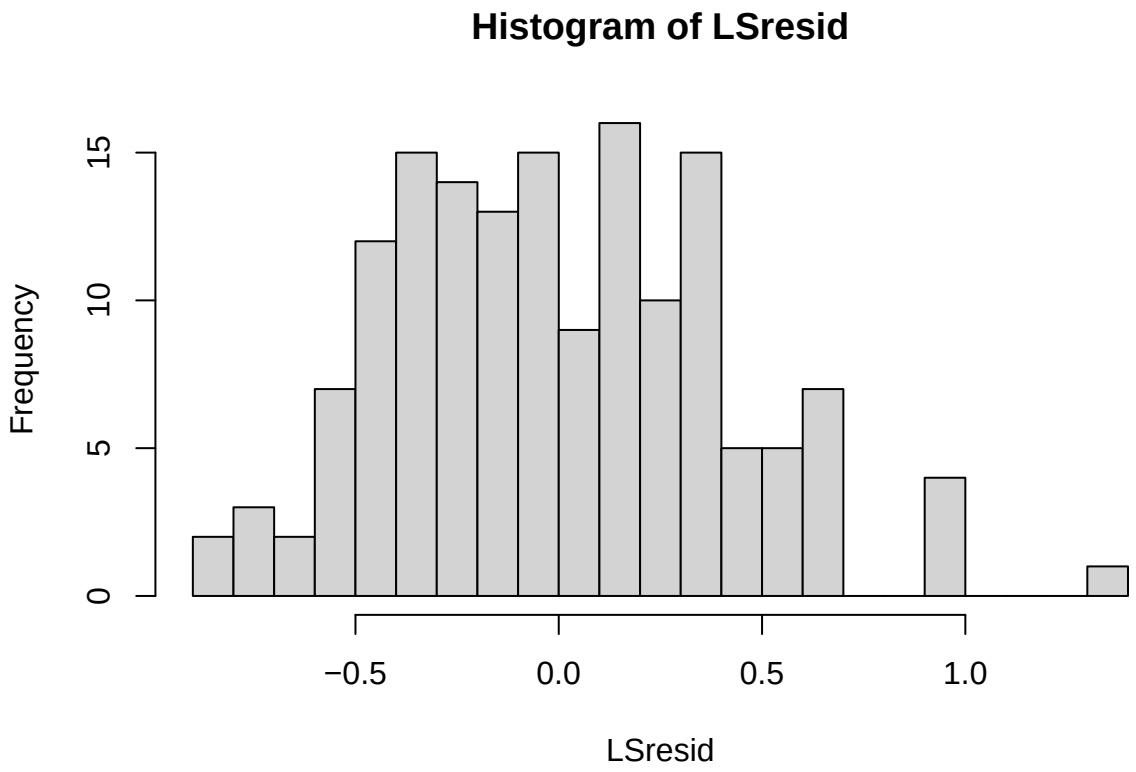
We continue the analysis using functions from the `geoR` library, for which a `geodata` data type is required. There are 155 observations (sampling locations)

```
zmat <- matrix(cbind(meuse$x,meuse$y,log(meuse$zinc)),
                 ncol=3,nrow=155,byrow=F)
geozinc <- as.geodata(zmat,coords.col=c(1,2),data.col=c(3))
```

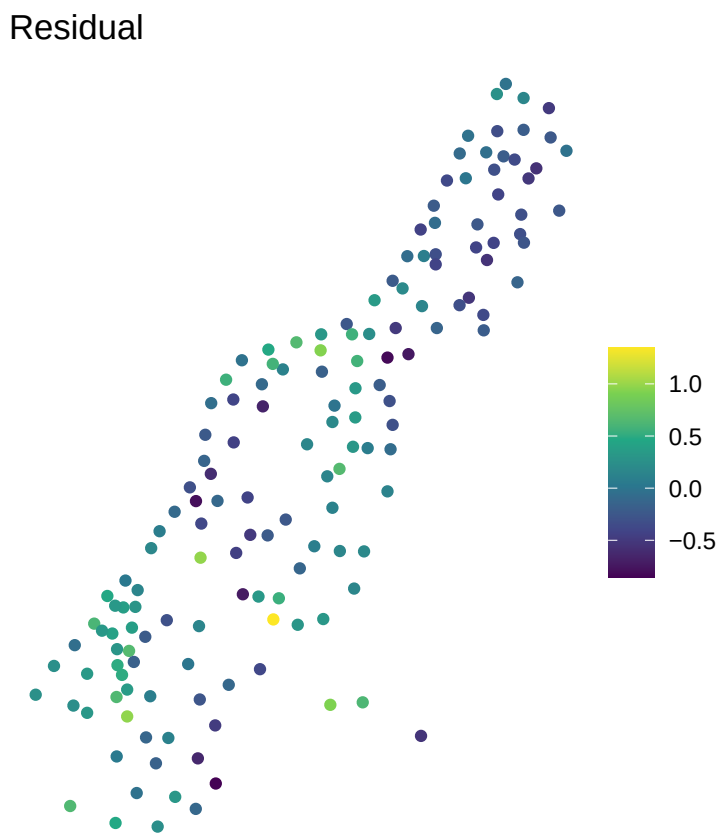
log(zinc) Variogram

First we will be assuming a spatial model on the residuals with `ffreq` and `sdist` in the model. Fit this initial linear model, and view the residuals by histogram and map.

```
LSmod <- lm(log(meuse$zinc)~sqrt(meuse$dist)+as.factor(meuse$ffreq))
LSresid <- residuals(LSmod,type="pearson")
hist(LSresid,nclass=25)
```




```
m.sf$resid <- LSresid
ggplot() + geom_sf(data = m.sf[,"resid"], aes(color = resid)) +
  theme_void() + scale_color_viridis_c() + labs(title = "Residual", color=NULL)
```



Moran's I

We can calculate Moran's I to see the strength of spatial dependence in the residuals. Our results show that we can reject the null hypothesis that there is zero spatial autocorrelation present in `LSresid` at the $\alpha = 0.05$ level.

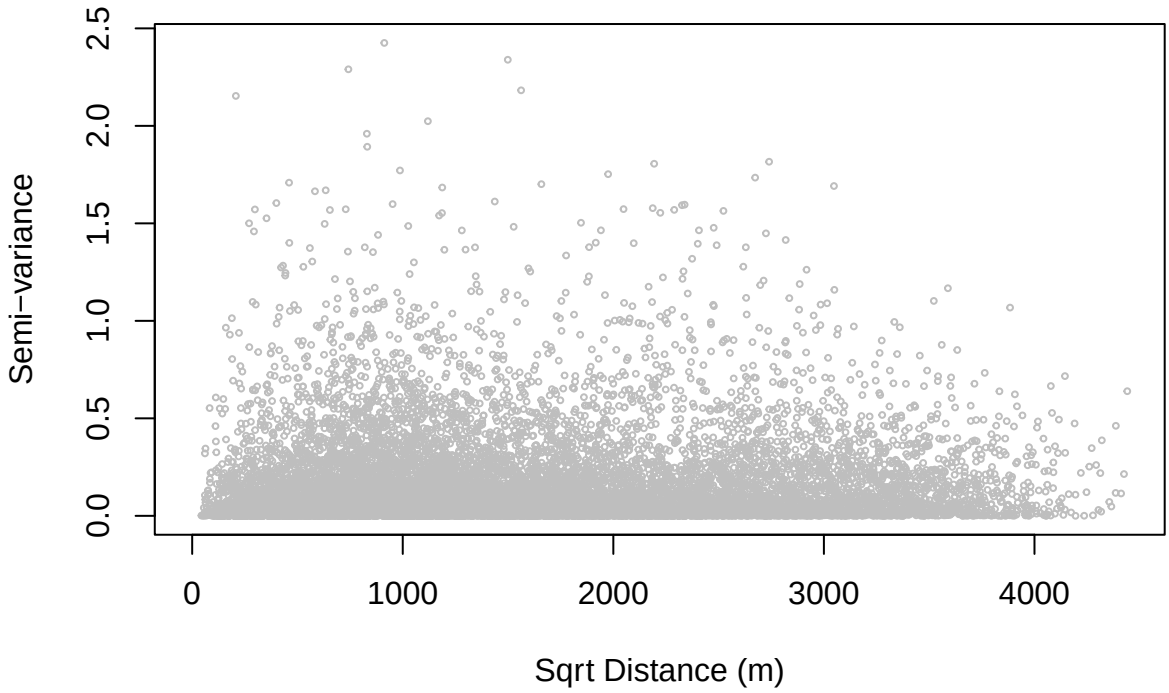
```
library(ape)
dists <- as.matrix(dist(cbind(geozinc[1]$coords[, 1], geozinc[1]$coords[, 2])))
dists.inv <- 1/dists
diag(dists.inv) <- 0
dists.inv[1:5, 1:5]
##             1           2           3           4           5
## 1 0.000000000 0.014116748 0.008414063 0.003857440 0.002729898
## 2 0.014116748 0.000000000 0.007063831 0.003535423 0.002757553
## 3 0.008414063 0.007063831 0.000000000 0.006984644 0.003983684
## 4 0.003857440 0.003535423 0.006984644 0.000000000 0.006482446
## 5 0.002729898 0.002757553 0.003983684 0.006482446 0.000000000

Moran.I(LSresid, dists.inv)
## $observed
## [1] 0.1210698
##
## $expected
## [1] -0.006493506
##
## $sd
```

```
## [1] 0.01060145
##
## $p.value
## [1] 0
```

Variogram Cloud

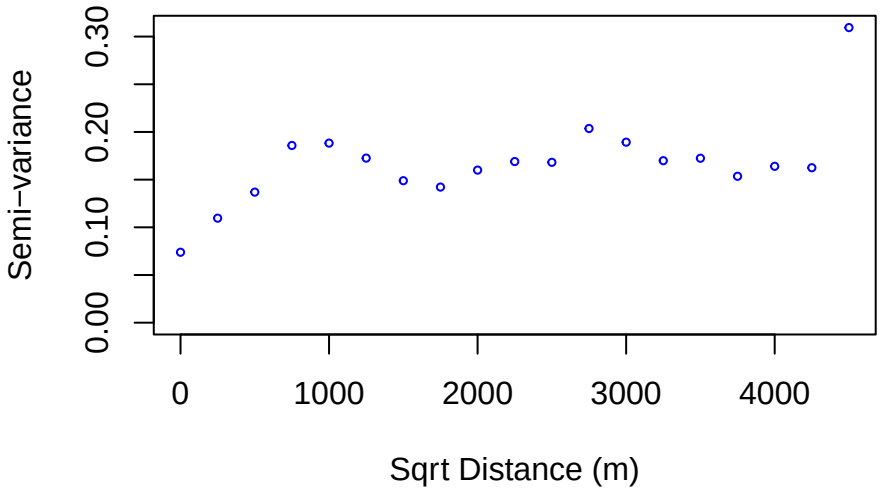
Variogram cloud for log zinc residuals, after taking out `ffreq` and `dist`.

```
cloudzinc <- variog(geozinc,option="cloud", trend=~sqrt(meuse$dist)+as.factor(meuse$ffreq))
## variog: computing omnidirectional variogram
plot(cloudzinc,ylab="Semi-variance",xlab="Sqrt Distance (m)",col="grey",cex=.4)
```



Binned Variogram

```
binzinc <- variog(geozinc,uvec=seq(0,5000,250),
           trend=~sqrt(meuse$dist)+as.factor(meuse$ffreq))
## variog: computing omnidirectional variogram
plot(binzinc,ylab="Semi-variance",xlab="Sqrt Distance (m)",cex=.5,col="blue")
```



Maximum likelihood for log(zinc)

We fit with the Matern covariance model with

$$\rho(h) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{h}{\phi}\right),$$

where h is the distance between 2 points and we take $\kappa = 0.5$ so that ϕ is the only estimated parameter. The other parameters estimated are τ^2, the nugget, and σ^2, the spatial variance. The practical range reported is the distance at which the correlations drop to 0.05, and is a function of ϕ.

We suppress the output from the call.

```
mlfit <- likfit(geozinc,cov.model="matern",ini=c(.2,224),trend=~sqrt(meuse$dist)+as.factor(meuse$ffreq)
```

We examine the results, specifically point estimates and standard errors.

```
mlfit$parameters.summary
##            status   values
## beta0   estimated   7.0712
## beta1   estimated  -2.1208
## beta2   estimated  -0.5154
## beta3   estimated  -0.5266
## tausq   estimated   0.0372
## sigmasq estimated   0.1316
## phi     estimated 300.5381
## kappa       fixed   0.5000
## psiA        fixed   0.0000
## psiR        fixed   1.0000
## lambda      fixed   1.0000
for (i in 1:3){
cat(cbind(mlfit$beta[i],sqrt(mlfit$beta.var[i,i])),"\n")
}
## 7.071249 0.1326244
## -2.120796 0.2365602
## -0.5153624 0.06780136
mlfit$practicalRange
## [1] 900.3318
```

Note that the standard errors change from the least squares fit.

Restricted maximum likelihood for log(zinc)

Restricted MLE is preferable in general for estimation when there are variance parameters.

```
remlfit <- likfit(geozinc,cov.model="matern",ini=c(.55,224),lik.method="RML",
         trend=~sqrt(meuse$dist)+as.factor(meuse$ffreq))
```

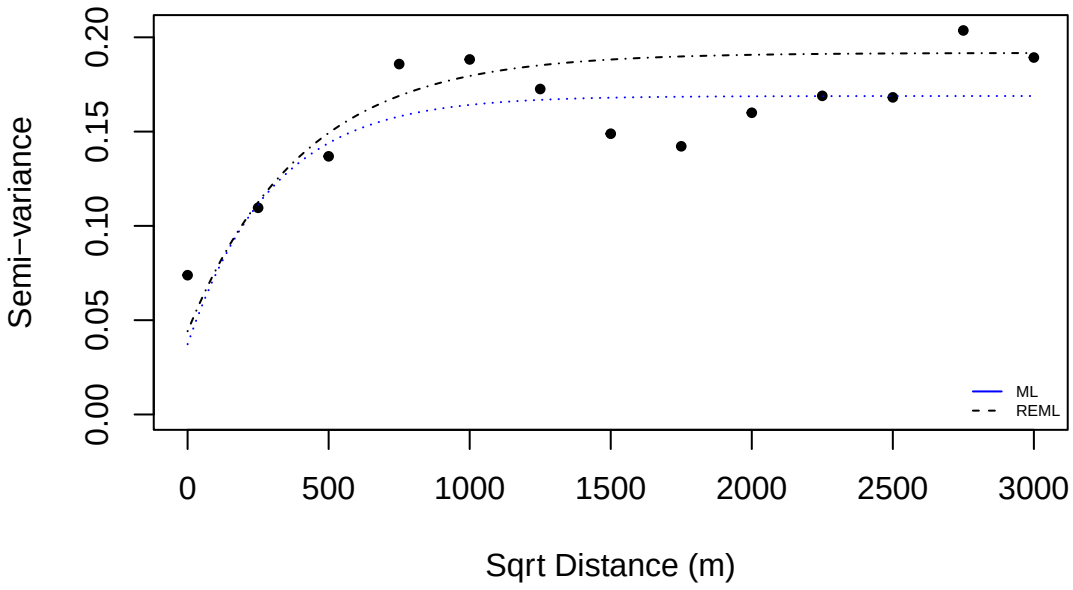
The results show slight differences from ML.

```
remlfit$parameters.summary
##            status   values
## beta0   estimated   7.0823
## beta1   estimated  -2.1109
## beta2   estimated  -0.5280
## beta3   estimated  -0.5455
## tausq   estimated   0.0442
## sigmasq estimated   0.1476
## phi     estimated 401.4313
## kappa       fixed   0.5000
## psiA        fixed   0.0000
## psiR        fixed   1.0000
## lambda      fixed   1.0000
remlfit$practicalRange
## [1] 1202.581
```

```
for (i in 1:3){
cat(cbind(mlfit$beta[i],sqrt(mlfit$beta.var[i,i]),remlfit$beta[i],sqrt(remlfit$beta.var[i,i])),"\n")
}
## 7.071249 0.1326244 7.082279 0.1527493
## -2.120796 0.2365602 -2.110917 0.2515499
## -0.5153624 0.06780136 -0.5280202 0.06913553
```

Comparison of estimates

```
plot(binzinc,max.dist=3000,xlab="Sqrt Distance (m)",ylab="Semi-variance",pch=19,cex=.6)
lines(mlfit,max.dist=3000,lty=3,col="blue")
lines(remlfit,max.dist=3000,lty=4,col="black")
legend("bottomright",legend=c("ML","REML"),
       lty=c(1,2),bty="n",col=c("blue","black"),cex=0.5)
```



Prediction for log(zinc) by Kriging

We now fit a linear model in distance and elevation to log(zinc). We then form a `geodata` object with the residuals as the response.

```
lmfit <- lm(geozinc$data ~ sqrt(meuse$dist) + as.factor(meuse$ffreq))
lmfit
##
## Call:
## lm(formula = geozinc$data ~ sqrt(meuse$dist) + as.factor(meuse$ffreq))
##
## Coefficients:
##             (Intercept)         sqrt(meuse$dist)  as.factor(meuse$ffreq)2
##                  7.0299                  -2.2660                  -0.3605
## as.factor(meuse$ffreq)3
##                 -0.3167
detrend <- as.geodata(cbind(geozinc$coords,lmfit$residuals))
```

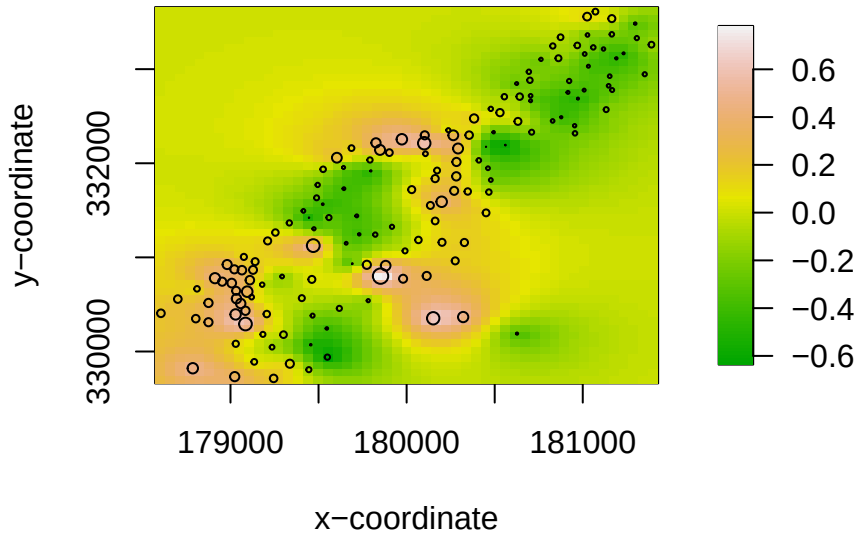
Finally, we can obtain spatial predictions on a grid, using the parameter estimates from the ML fit.

```
pred.grid <- expand.grid(seq(178600,181400,l=51),
             seq(329700,333620,l=51))
kc <- krige.conv(detrend,loc=pred.grid,
      krige=krige.control(obj.m=mlfit))
## krige.conv: model with constant mean
## krige.conv: Kriging performed using global neighbourhood
```

To view the results produce an image plot of the predictions, with the data superimposed.

```
library(fields)
image.plot(x=pred.grid[["Var1"]][1:51],y=unique(pred.grid[["Var2"]]),
           z=matrix(kc$predict,nrow=51,ncol=51),col=terrain.colors(100),xlab="x-coordinate",ylab="y-coo
```

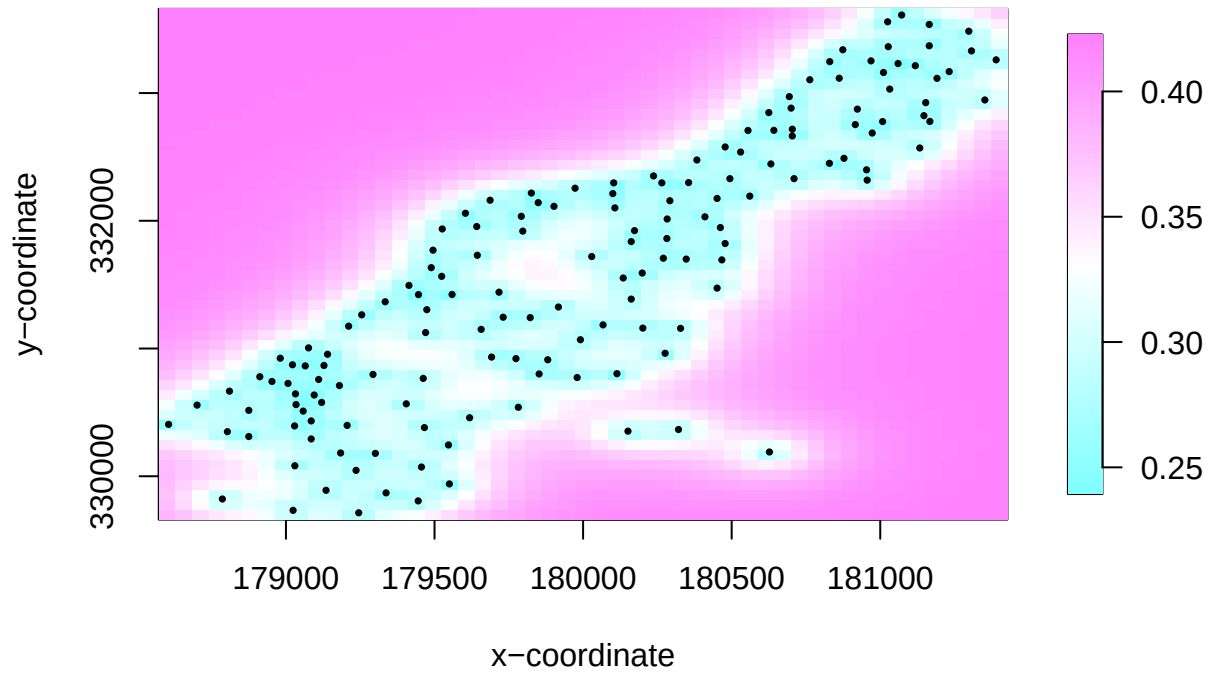
```
symbols(detrend$coords[,1],detrend$coords[,2],
  circles=(detrend$data-min(detrend$data))/1,add=T,inches=0.04)
```



Standard deviations of prediction for log(zinc)

We now plot the Kriging standard deviations of the predictions.

```
image.plot(x=pred.grid[["Var1"]][1:51],y=unique(pred.grid[["Var2"]]),
           z=matrix(sqrt(kc$krige.var),nrow=51,ncol=51),
           col=cm.colors(100),xlab="x-coordinate",ylab="y-coordinate")
points(detrend$coords[,1],detrend$coords[,2],cex=.5,pch=16)
```



The standard deviation is smallest close to the datapoints, as expected.

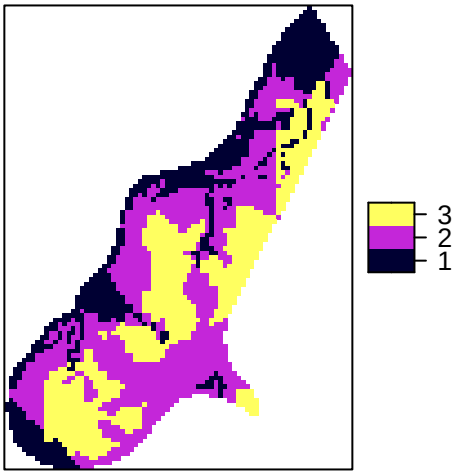
Another example, using categorical flooding frequency covariate

Now, we'll repeat, but use categorical flooding frequency and distance to the Meuse River as covariates. These two covariates are available for a grid, in the `meuse.grid` data object. This means we can predict log(Zinc) for a grid, using the sum of the linear model fit and the interpolated spatial field.

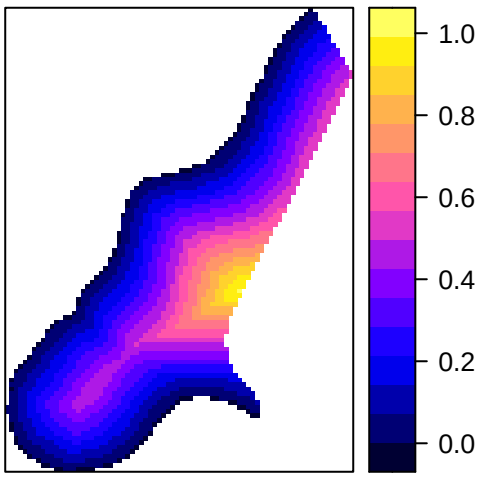
First take a look at the grid covariates.

```
library(sp)
data(meuse.grid)
coordinates(meuse.grid) = ~x+y
proj4string(meuse.grid) <- CRS("+init=epsg:28992")
gridded(meuse.grid) = TRUE
```

```
spplot(meuse.grid["ffreq"])
```



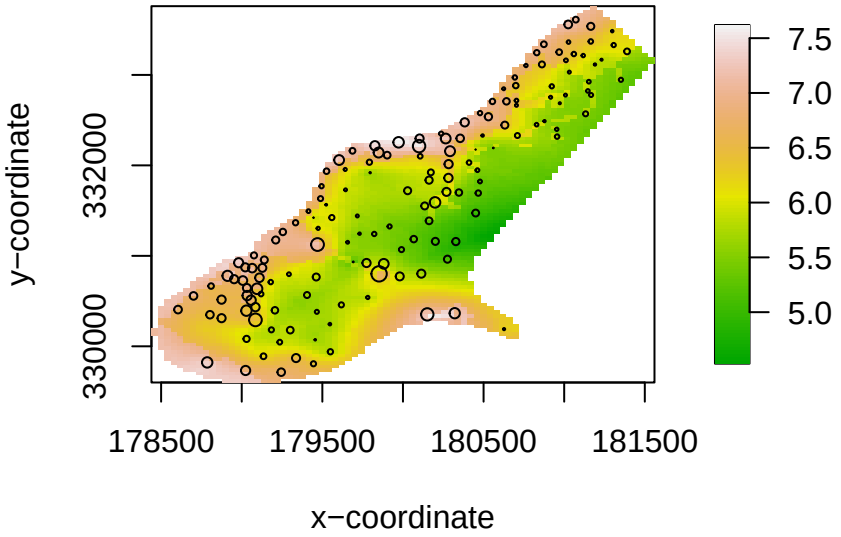

```
spplot(meuse.grid["dist"])
```



Now proceed with Kriging for log(Zinc).

```
# (label objects associated with this version with a "3")
lmdata <- data.frame(logzinc=geozinc$data, dist=sqrt(meuse$dist), ffreq=as.factor(meuse$ffreq))
lmfit3 <- lm(logzinc~dist+ffreq, data=lmdata)
detrend3 <- as.geodata(cbind(geozinc$coords,lmfit3$residuals))
mlfit3 <- likfit(detrend3,ini=c(.2,224))
## ------------------------------------------------------------------
## likfit: likelihood maximisation using the function optim.
## likfit: Use control() to pass additional
##          arguments for the maximisation function.
##         For further details see documentation for optim.
## likfit: It is highly advisable to run this function several
##         times with different initial values for the parameters.
## likfit: WARNING: This step can be time demanding!
## ------------------------------------------------------------------
## likfit: end of numerical maximisation.

# get meuse.grid and expand to all combinations
data(meuse.grid)
pred.grid3 <- expand.grid(sort(unique(meuse.grid$x)),
                          sort(unique(meuse.grid$y)))

# kriging
kc3 <- krige.conv(detrend3,loc=pred.grid3,
                  krige=krige.control(obj.m=mlfit3))
## krige.conv: model with constant mean
## krige.conv: Kriging performed using global neighbourhood

# merge linear predictions onto grid
pred_lm <- meuse.grid[, c("x", "y", "dist", "ffreq")]
pred_lm$pred_lm <- predict(lmfit3, newdata = pred_lm)
pred_resid <- data.frame(pred_resid = kc3$predict, x = pred.grid3$Var1,
                         y = pred.grid3$Var2)
pred_logzinc <- merge(pred_resid, pred_lm, by = c("x","y"), all=T)
pred_logzinc$pred_logzinc <- pred_logzinc$pred_lm + pred_logzinc$pred_resid
pred_logzinc <- pred_logzinc %>% arrange(y,x)
```

```
# plot prediction
image.plot(x=unique(pred.grid3[["Var1"]]),y=unique(pred.grid3[["Var2"]]),
           z=matrix(pred_logzinc$pred_logzinc,nrow=78,ncol=104),col=terrain.colors(100),
           xlab="x-coordinate",ylab="y-coordinate")
symbols(detrend3$coords[,1],detrend3$coords[,2],
        circles=(detrend3$data-min(detrend3$data))/1,add=T,inches=0.04)
```


y−coordinate
330000
332000
178500
179500
180500
181500
x−coordinate
7.5
7.0
6.5
6.0
5.5
5.0